PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

ESCUELA DE SISTEMAS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS Y COMPUTACIÓN

"ANÁLISIS COMPARATIVO DE HERRAMIENTAS OPEN SOURCE PARA DATA MINING SOBRE DATOS PÚBLICOS DEL MINISTERIO DE EDUCACIÓN DE LA REPÚBLICA DEL ECUADOR"

AUTOR: SERGIO DAVID PÁEZ JUKA

QUITO, JUNIO 2019

Índice de Contenidos

1. Introducción a la Minería de Datos 1
1.1 Definición y objetivos de Minería de Datos 1
1.2 Orígenes 4
1.3 Ciencias y Campos Relacionados con la Minería de Datos 8
1.4 Herramientas FLOSS o FOSS 11
2. Técnicas de Minería de Datos 13
2.1 Descripción del Proceso General de un proyecto de Minería de Datos 13
2.1.1 Definición del Problema 13
2.1.2 Exploración de los datos 14
2.1.3 Preparación de datos 15
2.1.4 Modelado 16
2.1.5 Evaluación 17
2.1.6 Lanzamiento 17
2.2 Aprendizaje Supervisado y No supervisado 18
2.2.1 Aprendizaje Supervisado 18
2.2.2 Aprendizaje no Supervisado 21
2.3 Casos de Aprendizaje Supervisado 25
2.3.1 Regresión Lineal - Definición y aclaraciones previas 25
2.3.2 Modelos Lineales y Mínimos Cuadrados 26
2.3.3. Método de los vecinos más cercanos (k-vecinos más próximos) 29
2.3.3.1 Ejemplos 29
2.3.4 Modelos de Regresión Lineal 30
2.3.5 Clasificación 32
2.5 Análisis Clúster 33

2.5.1 Definición 33
2.5.2 Matrices de Proximidad 35
2.5.3 Algoritmos Clústering 35
2.6 Análisis del Marco Teórico Incluido 36
2.6.1 El Aprendizaje de Máquina en Minería de Datos. 36
2.6.2 La importancia de los distintos métodos tratados. 37
3. Obtención de Datos y Elección de Herramientas. 39
3.1 Definición de los repositorios de datos a usar desde el Ministerio de Educación. . 39
3.2 Análisis previo de los datos elegidos. 41
3.2.1 Descripción de la estructura de los archivos 41
3.2.2 Preparación de los Datos 42
3.2.3 Gráficas Preliminares 45
3.3 Herramientas de Minería de Datos elegidas 48
3.3.1 R y R Studio 48
3.3.2 Orange 50
3.3.3 Motivos para la selección de Herramientas 51
4. Aplicación de Minería de Datos 52
4.1 Introducción a las Herramientas Elegidas. 52
4.1.1 Minería de Datos con R y RStudio 52
4.1.2 Minería de Datos con Orange 53
4.1.3 Aproximación para el presente trabajo 54
4.1.4 Aclaración con respecto a las métricas 55
4.1.5 Objetivos y la generación de predicciones 55
4.2 Aplicación de Minería de Datos usando R 56
4.2.1 Preprocesamiento 56

4.2.2 Visualización de los Datos 58
4.2.3 Minería de Datos (Modelado y Evaluación de Datos) 62
4.3 Aplicación de Minería de Datos usando Orange 71
4.3.2 Preprocesamiento 75
4.3.3 Visualización de Datos 77
4.3.4 Minería de Datos (Modelado y Evaluación de Datos) 80
5. Análisis Comparativo de Herramientas 91
5.1 Evaluación previa de resultados obtenidos con las herramientas seleccionadas 91
5.1.1 Comparación Previa del trabajo con las herramientas seleccionadas. 91
5.1.2 Resultados previos obtenidos de cada hipótesis. 94
5.2 Establecimiento de parámetros para evaluar resultados. 99
5.3 Determinación de tasas de error de cada resultado mediante validación cruzada. . 99
5.3.1 Métricas Hombres Promovidos Segundo de Básica 104
5.3.2 Métricas Mujeres Abandono Tercero de Bachillerato 106
5.3.3 Métricas Hombres No Promovidos Octavo de Básica 109
5.4 Cuadro Comparativo de Herramientas 113
6. Conclusiones y Recomendaciones 114
6.1 Determinación de ventajas y desventajas de las herramientas seleccionadas. 114
6.1.1 Ventajas del trabajo en RStudio 114
6.1.2 Desventajas del trabajo en RStudio 115
6.1.3 Ventajas del trabajo en Orange 115
6.1.4 Desventajas del trabajo en Orange 117
6.2 Ética en la Minería de Datos 119
6.2 Conclusiones de la aplicación de Minería de Datos sobre los datos del Ministerio de Educación 121

6.3 Recomendaciones para futuras aplicaciones de Minería de Datos sobre fuentes similares 122

7. Bibliografía 124

Índice de Tablas

Tabla 1. R y RStudio..... 49
Tabla 2. Orange..... 50
Tabla 3- Capítulo 5 Métricas Hombres Promovidos Segundo de Básica – Versión 1 104
Tabla 4 - Capítulo 5 Métricas Hombres Promovidos Segundo Básica - Versión 2..... 104
Tabla 5 - Capítulo 5 Métricas Mujeres Abandono Tercero de Bachillerato – Versión 1 106
Tabla 6 - Capítulo 5 Métricas Mujeres Abandono Tercero de Bachillerato - Versión 2..... 107
Tabla 7 - Capítulo 5 Métricas Hombres No Promovidos Octavo de Básica - Versión 1 109
Tabla 8 - Capítulo 5 Métricas Hombres No Promovidos Octavo de Básica - Versión 2 110
Tabla 9 Cuadro Comparativo de Herramientas 113

Índice de Figuras

Figura 1. Proceso General del KDD. En Data Mining: Concepts and Techniques (p. 7), por Han, Kamber y Pei..... 3
Figura 2. Fases del Proceso de Minería de Datos. En IBM SPSS Modeler CRISP-DM Guide (p. 1), por IBM. 3
Figura 3. Historia de las colecciones de Datos. En Data Mining: Concepts and Techniques (p. 3), por Han, Kamber y Pei..... 7
Figura 4. Ejemplos de pares x,y con sus respectivas hipótesis. En Artificial Intelligence: A Modern Approach (p. 696), por Russell y Norvig. 20
Figura 5. Un ejemplo de clasificación en dos dimensiones, con una codificación de clases en una variable binaria (AZUL= 0, NARANJA=1) y luego, ajustadas por una regresión lineal. En The Elements of Statistical Learning (p. 13), por Hastie, Tibshirani y Friedman. 28
Figura 6. El mismo ejemplo de clasificación que en la Figura 5. El ajuste está realizado por el método de los vecinos próximos (15-vecinos próximos). En The Elements of Statistical Learning (p. 15), por Hastie, Tibshirani y Friedman (2009). 29

Figura 7. El mismo ejemplo de clasificación que en la Figura 5. El ajuste está realizado por el método de los vecinos próximos (1-vecino próximo). En The Elements of Statistical Learning (p. 16), por Hastie, Tibshirani y Friedman (2009). 30
Figura 8. Datos simulados en el plano, agrupados en tres clases (representadas en color naranja, azul y verde) por el algoritmo de Clustering "K-medias". En The Elements of Statistical Learning (p. 502), por Hastie, Tibshirani y Friedman (2009). 34
Figura 9. Conjunto de Datos de Instituciones Educativas en el Ecuador, por Ministerio de Educación..... 40
Figura 10. Revisión de resultados para encontrar posibles columnas en cero. (Elaboración Propia)..... 44
Figura 11. Evolución del número total de docentes mujeres en el Ecuador (Elaboración Propia)45
Figura 12. Evolución del número total de docentes hombres en el Ecuador. (Elaboración Propia) 46
Figura 13. Evolución del número de administrativos mujeres de instituciones educativas en el Ecuador. (Elaboración Propia)..... 46
Figura 14. Evolución del número de administrativos hombres de instituciones educativas en el Ecuador (Elaboración Propia)..... 47
Figura 15. Evolución del número de estudiantes mujeres de Educación Básica y Bachillerato en el Ecuador (Elaboración Propia)..... 47
Figura 16. Evolución del número de estudiantes hombres de Educación Básica y Bachillerato en el Ecuador (Elaboración Propia)..... 48
Figura 17. Importación de datos en RStudio (Elaboración Propia)..... 56
Figura 18. Visualización de Datos Importados en RStudio (Elaboración Propia) 57
Figura 19. Ventana de instalación de paquetes adicionales en RStudio (Elaboración Propia)..... 57
Figura 20. Importación de la biblioteca "dplyr" (Elaboración Propia)..... 58
Figura 21. Selección de datos usando la biblioteca "dplyr" (Elaboración Propia) 58
Figura 22. Usando la biblioteca "ggplot2" (Elaboración Propia) 59
Figura 23. Generación de un histograma del número de docentes mujeres usando la biblioteca "ggplot2" en RStudio. (Elaboración Propia)..... 59
Figura 24. Mensajes de advertencia en RStudio al usar la biblioteca "ggplot2" (Elaboración Propia)..... 60

Figura 25. Histograma con datos de docentes mujeres obtenido mediante la biblioteca "ggplot2" en RStudio (Elaboración Propia) 60
Figura 26. Relación entre los Docentes Femeninos y los estudiantes varones promovidos del segundo año de educación básica (Elaboración Propia) 61
Figura 27. Gráfico que relaciona las variables de la Figura 26 más la variable de los Administrativos Mujeres. (Elaboración Propia) 61
Figura 28. Gráfico resultante del código descrito en la Figura 27. (Elaboración Propia) 62
Figura 29. Función para separar los datos en base a una variable aleatoriamente elegida (Elaboración Propia) 63
Figura 30. Separación de los datos en dos subconjuntos “train” y “test” mediante las etiquetas de TRUE o FALSE establecidas previamente. (Elaboración Propia) 63
Figura 31. Aplicación de la función lm con la variable HombresPromovidosSegundoBasica como variable dependiente. (Elaboración Propia) 64
Figura 32. Predicción de nuevos valores usando la función "predict" en RStudio. (Elaboración Propia) 64
Figura 33. Comparación de los resultados actuales y los predichos de los datos obtenidos en RStudio. (Elaboración Propia) 64
Figura 34. Resumen de la tabla de comparación de resultados (actuales y predichos) de lo obtenido en RStudio. (Elaboración Propia) 65
Figura 35. Procedimiento en forma de comandos para mostrar el error de los datos actuales y los predichos en RStudio. (Elaboración Propia) 65
Figura 36. Resumen de la tabla de comparación de resultados con su error respectivo del trabajo realizado en RStudio. (Elaboración Propia) 66
Figura 37. Obteniendo el error cuadrático medio en base a los valores obtenidos previamente en el trabajo con RStudio. (Elaboración Propia) 66
Figura 38. Resumen del modelo realizado en RStudio. (Elaboración Propia) 66
Figura 39. Influencia de variables en el modelo realizado en RStudio. (Elaboración Propia) 67
Figura 40. Métricas correspondientes al modelo realizado en RStudio, entre ellos, el “R Cuadrado Ajustado”. (Elaboración Propia) 68
Figura 41. Resultados de una nueva versión del modelo realizado en RStudio reduciendo las variables solamente a las relacionadas con la hipótesis. (Elaboración Propia) 69

Figura 42. Resultados del Primer Modelo en RStudio con la variable "MujeresAbandonoTerceroBach" como dependiente. (Elaboración Propia) 69
Figura 43. Modelo en RStudio actualizado con la variable Mujeres Abandono Tercero de Bachillerato donde se mantiene el valor del R Cuadrado Ajustado. (Elaboración Propia) 70
Figura 44. Resumen del segundo modelo aplicado con variable dependiente como "HombresNOAprovadosOctavoBásica" y eliminando las variables que en el primer modelo se consideraron como poco influyentes. (Elaboración Propia) 71
Figura 45.Interfaz de Orange y opciones de la sección "Data". (Elaboración Propia) 72
Figura 46. Interfaz de Orange. Sección "Visualize" y sus diferentes widgets. (Elaboración Propia) 73
Figura 47. Interfaz de Orange. Sección "Model" y sus distintos widgets. (Elaboración Propia) . 73
Figura 48. Interfaz de Orange. Sección "Evaluate" y sus diferentes widgets. (Elaboración Propia) 74
Figura 49. Interfaz de Orange. Sección "Unsupervised" y sus diferentes widgets .(Elaboración Propia) 74
Figura 50. Ventana de Orange con la carga de datos; la variable Período se encuentra seleccionada como variable objetivo. 75
Figura 51. Selección de columnas para el modelo (a la izquierda se muestran las columnas desafectadas) en Orange. (Elaboración Propia) 76
Figura 52.Uniendo dos widgets en Orange. (Elaboración Propia) 77
Figura 53. Concatenación de widgets de diagramas elegidos en Orange para obtener visualización de los datos. (Elaboración Propia) 77
Figura 54. Distribución de la variable Docentes Femenino agrupada por la variable Tipo Educación usando Orange. (Elaboración Propia) 78
Figura 55. Ejemplo de Diagrama de Dispersión en Orange. (Elaboración Propia) 79
Figura 56. Diagrama Mosaico en Orange. (Elaboración Propia) 80
Figura 57. Función write en R para guardar los archivos de prueba y entrenamiento y usarlos en Orange. (Elaboración Propia) 81
Figura 58. Carga de datos y selección de los mismos en Orange. (Elaboración Propia) 81
Figura 59. Generación de predicciones. Se muestran los flujos de trabajo de los datos de entrenamiento y prueba en Orange. (Elaboración Propia) 82

Figura 60. Predicciones obtenidas con el proceso realizado en Orange. (Elaboración Propia).... 83
Figura 61. Modelo de Minería de Datos usando Regresión Lineal en Orange. (Elaboración Propia)..... 83
Figura 62. Resultados del Modelo de Regresión Lineal en Orange mostrado anteriormente. (Elaboración Propia) 84
Figura 63. Modelo usando Redes Neuronales en Orange. (Elaboración Propia) 85
Figura 64. Resultados del Modelo con Redes Neuronales en Orange. (Elaboración Propia) 85
Figura 65. Modelo reducido sin la necesidad de particionar los datos en Orange. (Elaboración Propia)..... 85
Figura 66. Ranking de las variables según el algoritmo RReliefF con la variable Hombres Promovidos Segundo Básica como variable objetivo en Orange. (Elaboración Propia)..... 86
Figura 67. Segunda Versión del Modelo con la variable Hombres Promovidos Segundo Básica como objetivo en Orange. (Elaboración Propia)..... 87
Figura 68. Primer modelo realizado con la variable Mujeres Abandono Tercero Bach como objetivo en Orange. (Elaboración Propia) 87
Figura 69. Resultados modelo de Mujeres Abandono Tercero Bach como variable objetivo usando Random Forest en Orange. (Elaboración Propia)..... 88
Figura 70. Ranking de las variables según el algoritmo RReliefF con la variable Mujeres Abandono Tercero Bach como objetivo y usando el método Random Forest en el modelo en Orange. (Elaboración Propia) 89
Figura 71. Modelo simplificado usando varios métodos en Orange. (Elaboración Propia)..... 89
Figura 72. Resultados del modelo realizado en Orange usando tres tipos distintos de métodos con la variable Hombres NO Aprobados Octavo Básica como objetivo. (Elaboración Propia) 90
Figura 73. Ranking de las variables según el algoritmo RReliefF con la variable Hombres NO Aprobados Octavo Básica como objetivo en Orange. (Elaboración Propia)..... 90
Figura 74. Trabajo con R Studio, donde se muestran sus respectivas subsecciones. (Elaboración Propia)..... 92
Figura 75. Trabajo con Orange, donde se muestra la concatenación de widgets para construir un modelo. (Elaboración Propia)..... 93

Figura 76. Resultados de uno de los modelos con la variable Hombres Promovidos Segundo Básica como objetivo en RStudio, donde se aprecia la relevancia de la variable Docentes.Femenino en el modelo. (Elaboración Propia)..... 95
Figura 77. Resultados de uno de los modelos con la variable Hombres Promovidos Segundo Básica como objetivo en ORANGE, donde se aprecia la relevancia de las variables de hombres promovidos de los distintos años de educación en Primaria. (Elaboración Propia) 95
Figura 78. Comparación entre dos períodos lectivos del número de estudiantes mujeres que abandonan sus estudios al cursar el Tercero de Bachillerato. (Elaboración Propia) 96
Figura 79. Resultados de uno de los modelos con la variable Hombres NO Promovidos Octavo Básica como objetivo en R Studio, donde se aprecia la relevancia de la variable Sostenimiento en el modelo. (Elaboración Propia) 98
Figura 80. Resultados de uno de los modelos con la variable Hombres NO Promovidos Octavo Básica como objetivo en ORANGE, donde se aprecia la relevancia de la variable Sostenimiento. (Elaboración Propia) 98
Figura 81. Tabla con las predicciones y errores obtenidas con uno de los modelos del período lectivo 2016-2017 usando RStudio. (Elaboración Propia) 100
Figura 82. Extensión del flujo de trabajo en cada modelo realizado en Orange para poder exportar las predicciones resultantes. (Elaboración Propia) 101
Figura 83. Selección de variables en Orange para su exportación en un archivo de extensión csv. (Elaboración Propia) 101
Figura 84. Archivo de extensión csv modificado con los errores correspondientes calculados y las demás métricas copiadas de los modelos de Orange. (Elaboración Propia) 102

1. Introducción a la Minería de Datos

1.1 Definición y objetivos de Minería de Datos

Diariamente grandes cantidades de datos son recolectadas en muchas áreas de la actividad humana, y la necesidad que tenemos de encontrar la utilidad de dichos datos hacen que el análisis de éstos sea importante. Hablamos de grandes cantidades de datos en muchas situaciones como investigaciones, negocios, entretenimiento, etcétera.

Tal cantidad de datos es el resultado de la computarización de nuestro mundo, que ha supuesto la creación de herramientas que facilitan el almacenamiento y gestión de los datos. Lo anterior se puede comprobar mirando a las muchas organizaciones que poseen una gran colección de datos que pasan por reseñas de usuarios, tipos de productos, todos los datos involucrados en transacciones, historial de ventas, etcétera. Empresas como supermercados se encargan, a la semana, de realizar millones de transacciones que son almacenadas sistemas computarizados; asimismo, los científicos, entre experimentos, observaciones, mediciones, etcétera, se encargan de almacenar muchos datos puros. Las empresas de telecomunicaciones viven a diario con un tráfico considerable proveniente de las intercomunicaciones de las personas. Las billones de búsquedas diarias en la web, los miles de videos almacenados en línea, la información de pacientes de la industria de la medicina, y muchos otros componen una lista de fuentes de datos que no para de crecer actualmente.

Las organizaciones comprendieron que los datos netos que recolectaban podían convertirse en un beneficio si se aplicaba alguna clase de proceso sobre ellos. Es justamente eso, lo que busca la minería de datos, encontrar un bien mayor tomando como base a los datos puros; la minería de datos convierte un gran conjunto de datos en conocimiento. El objetivo es, por lo tanto, encontrar patrones de un conjunto de datos, que nos lleven a levantar conocimiento que, usualmente, no puede ser encontrado estudiando los datos por separado, o también encontrar predicciones para en base a éstos, tener un mayor control sobre los datos.

A pesar de que previamente se estableció el objetivo central de la Minería de Datos, existen muchas definiciones válidas.

> Data Mining consiste en un conjunto de metodologías estadísticas y computacionales que, junto a un enfoque desde las ciencias de la conducta, permite el

> análisis de datos y la elaboración de modelos matemáticos descriptivos y predictivos de la conducta del consumidor (Palma, Palma, & Pérez, 2009, pág. 43)

Podemos decir pues, que la Minería de Datos es un grupo de metodologías apoyadas en la Estadística y la Computación que permiten analizar datos para obtener conocimiento que entregue un valor agregado a los interesados en dichos datos.

Un objetivo derivado de la búsqueda de encontrarle utilidad a un conjunto de datos es el enfoque hacia la predicción de eventos. Como bien menciona Casas, Gironés, Minguillón, & Caihuelas (2017) la Física en la actualidad apunta a un estudio de un universo de eventos y no tanto de partículas, lo que hace que muchas ramas de la Ciencia se encarguen del entendimiento y/o predicción de los mencionados eventos; precisamente la minería de datos encuentra una de sus principales metas en este aspecto de pronóstico de situaciones, adelantarse a ciertos escenarios con el fin de obtener ventaja. Sin embargo, también es preciso anotar que la simple visualización de datos tiene una utilidad limitada que termina por insatisfacer la meta antes mencionada de la minería de datos; esto supone que detrás, existan procesos fundamentados correctamente para poder realizar análisis de los datos.

Tal y como mencionan Gibert, Ruiz, & José (2006) debido a los conceptos que se mencionaron, se tiende a confundir a la minería de datos como un sinónimo del proceso de descubrimiento de información de los datos (o KDD por sus siglas en inglés); mientras que otros ven a la minería de datos como un paso más dentro del proceso general de KDD. Por tanto, es necesario mencionar cada uno de los pasos de este proceso para obtener una nueva definición de Minería de Datos:

- Limpieza de datos: aquí se busca eliminar el ruido e inconsistencia de los datos, es decir, eliminar aquellos datos que, a simple vista, no aportarán nada para la construcción del conocimiento.
- Integración de datos: aquí se busca la combinación de varias fuentes de datos para simplificar el resto del proceso.
- Selección de datos: paso en donde se extrae de una base de datos, todo aquello que sea relevante para el análisis actual.
- Minería de datos: paso esencial que aplica distintos métodos para extraer patrones de los datos, o realiza predicciones de los mismos.

- Evaluación de resultados: luego de obtener patrones y/o predicciones es necesario identificar cuáles de ellos son útiles para el análisis actual y eliminar aquellos que no aporten valor.
- Presentación de Conocimiento: paso final donde se presenta a los usuarios, mediante técnicas de visualización y representación, el conocimiento que ha sido minado. (Han, Kamber, & Pei, 2012).

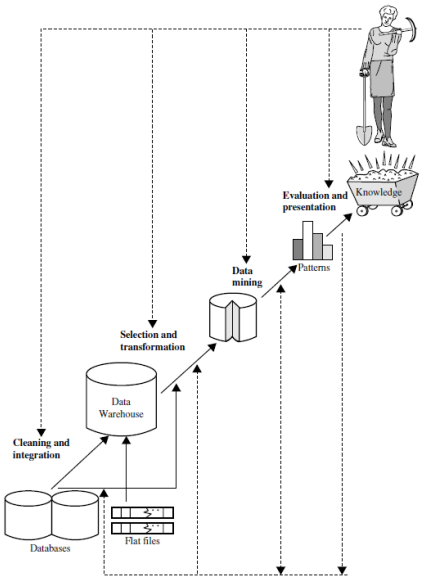


Figura 1. Proceso General del KDD. En Data Mining: Concepts and Techniques (p. 7), por Han, Kamber y Pei.

Existe un proceso más específico para la Minería de Datos, que se muestra a continuación:

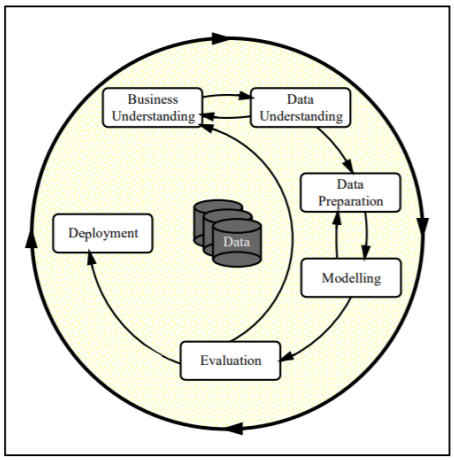


Figura 2. Fases del Proceso de Minería de Datos. En IBM SPSS Modeler CRISP-DM Guide (p. 1), por IBM.

Este proceso iterativo consta de 6 pasos:

1. Definición del problema o Entendimiento del Negocio
2. Exploración de los datos o Entendimiento de los Datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Lanzamiento

Cada uno de esos puntos serán abarcados a profundidad en posteriores capítulos.

1.2 Orígenes

Como mencionan Han, Kamber, & Pei (2012) la Minería de Datos encuentra su origen en el proceso natural de evolución de la Tecnología Informática y la industria del manejo de datos en general. Especialmente este último campo trajo avances fundamentales como la creación de bases de datos y su manejo dando como resultado la creación de términos como Almacén de Datos (Data Warehouse) y la misma Minería de Datos; conceptos que desde su inicio ya involucraban procesos avanzados de análisis de datos.

El mismo hecho de el desarrollo de colecciones de datos en conjunto con la creación de mecanismos, o bien gestores, de bases de datos sirvieron como base para el posterior funcionamiento de mecanismos efectivos de guardado de datos, así como su recuperación y visualización, que involucró toda una estructura de consultas y transacciones (que llegarían a priori a una estandarización con lenguajes como el Structured Query Language). Por lo que, métodos de análisis avanzado de datos sería el siguiente paso natural.

Desde la década de 1960, y hasta nuestros días, la manera en que hemos almacenado información ha evolucionado constantemente. El ritmo ha sido tan acelerado que hemos pasado en apenas cincuenta años, aproximadamente, de tener sistemas de procesamiento de archivos primitivos, a motores de bases de datos que permiten la creación de almacenes en poco tiempo y de manera gráfica para que la mayor cantidad de personas pueda realizar este proceso. En la actualidad contamos con muchas facilidades para manejar los datos, como las herramientas de modelado que permiten la construcción de diagramas para simplificar la visualización de elementos en una estructura relacional; la indexación y métodos de acceso a los datos; etcétera. Además, los usuarios de estos gestores de bases de datos han ganado flexibilidad en el manejo de los datos mediante la constante innovación de las interfaces de usuario, pero también la creación

de los ya mencionados lenguajes de consulta e inclusive las facilidades en el manejo de transacciones. (Han et al., 2012)

Sin embargo, todas esas características del manejo de los datos que fueron mencionadas son parte del avance de muchos años atrás. Luego del establecimiento de los sistemas de manejo de bases de datos, también llamados "gestores", el campo de las bases de datos se orientó a tres pilares fundamentales: los denominados "sistemas avanzados de bases de datos", el "data warehousing" y el "data mining". Estos sistemas avanzados nacieron como parte de la investigación de mediados de los años 80 y se centraron en incorporan funcionalidades como modelos de datos más eficientes como los orientados a objetos y los objeto-relacionales e inclusive modelos deductivos. En ese momento se dio el "boom" de los sistemas de propósito general que hacían uso de todas estas nuevas características y nacieron bases de datos espaciales, de multimedia, científicas, de ingeniería, de conocimiento, etcétera.

El análisis avanzado de datos surgió a finales de la década de los 80 en donde se tenía un mercado con hardware evolucionando constantemente, computadores potentes y asequibles, y sobre todo, equipos de recolección y almacenamiento de datos que potenció la industria de las bases de datos y la información, que a su vez provocó el incremento de bases de datos y repositorios de información disponibles para el manejo de transacciones, recuperación de información y análisis de datos. Uno de los tantos tipos de repositorios que surgieron fueron los denominados "Data Warehouse" o simplemente "Almacenes de Datos" que se caracteriza por ser un repositorio de múltiples y heterogéneas fuentes de datos, organizadas bajo un esquema unificado en un solo lugar para facilitar el manejo de la toma de decisiones. Estos almacenes de datos incluían técnicas ya citadas en el presente trabajo como la limpieza de datos y la integración de datos, además de procesamiento analítico en línea (OLAP por su siglas en inglés), que básicamente comprende técnicas de análisis con funcionalidades tales como realización de resúmenes, consolidación, agregación, así como la habilidad de "ver" la información desde distintos ángulos. (Han et al., 2012)

A pesar de que esta herramienta (OLAP) soporta análisis multidimensional y toma de decisiones, se necesitan herramientas de análisis de datos adicionales para llevar a cabo un análisis más profundo; por ejemplo, las herramientas de Minería de Datos proveen clasificación de datos, Clustering, detección de anomalías y visualización de cambios de los datos a lo largo del tiempo.

Poco después, durante la década de 1990, la World Wide Web (WWW) y las bases de datos orientadas a la web empezaron a dar sus primeros signos de vida. Las bases de información globales basadas en Internet, como la WWW y varios tipos de bases interconectadas y heterogéneas empezaron a surgir y tomaron un rol importante en la industria de la información. A pesar de que ya han transcurrido dos décadas, aproximadamente, de la globalización del Internet, el análisis efectivo y eficiente de datos provenientes de distintas fuentes mediante la integración de: recuperación de información, minería de datos y el uso de tecnologías de análisis de redes de información, sigue siendo una tarea desafiante hoy en día. En resumen, la abundancia de datos, junto con la necesidad de potentes herramientas de análisis de datos, se ha descrito como una situación rica en datos, pero pobre en información. La enorme cantidad de datos, de rápido crecimiento, recopilados y almacenados en grandes y numerosos repositorios de datos, ha superado con creces nuestra capacidad humana de comprensión sin usar herramientas poderosas para llevar a cabo estos análisis. (Han et al., 2012)

Como resultado, los datos recopilados en grandes almacenes de datos se convierten en "tumbas de datos", archivos de datos que rara vez se visitan. En consecuencia, las decisiones importantes a menudo se toman, no en base a los datos ricos en información almacenados en los repositorios, sino más bien en la intuición de aquel que esté a cargo de tomar las decisiones, simplemente porque dicha persona no tiene las herramientas necesarias para extraer el valioso conocimiento incorporado en las vastas cantidades de datos.

Se han realizado esfuerzos para desarrollar sistemas expertos y tecnologías basadas en el conocimiento, que generalmente dependen de los usuarios o expertos en el dominio para ingresar manualmente el conocimiento en las bases de conocimiento.

Desafortunadamente, el procedimiento de entrada de conocimiento manual es propenso a sesgos y errores y es extremadamente costoso y lento. La creciente brecha entre los datos y la información exige el desarrollo sistemático de herramientas de minería de datos que pueden convertir las tumbas de datos en "pepitas de oro" del conocimiento que aboguen por la creación de un mundo de datos más útil para todos. A continuación, se muestra un gráfico con los avances más importantes de las herramientas que trabajan con los datos, ejemplificando lo que se ha expuesto en esta sección; adicionalmente se incluyen varias otras herramientas que, no fueron mencionadas, pero que su aporte no deja de ser fundamental en todo este proceso de evolución de cómo el ser humano, mediante tecnología, manipula los datos para obtener beneficios:

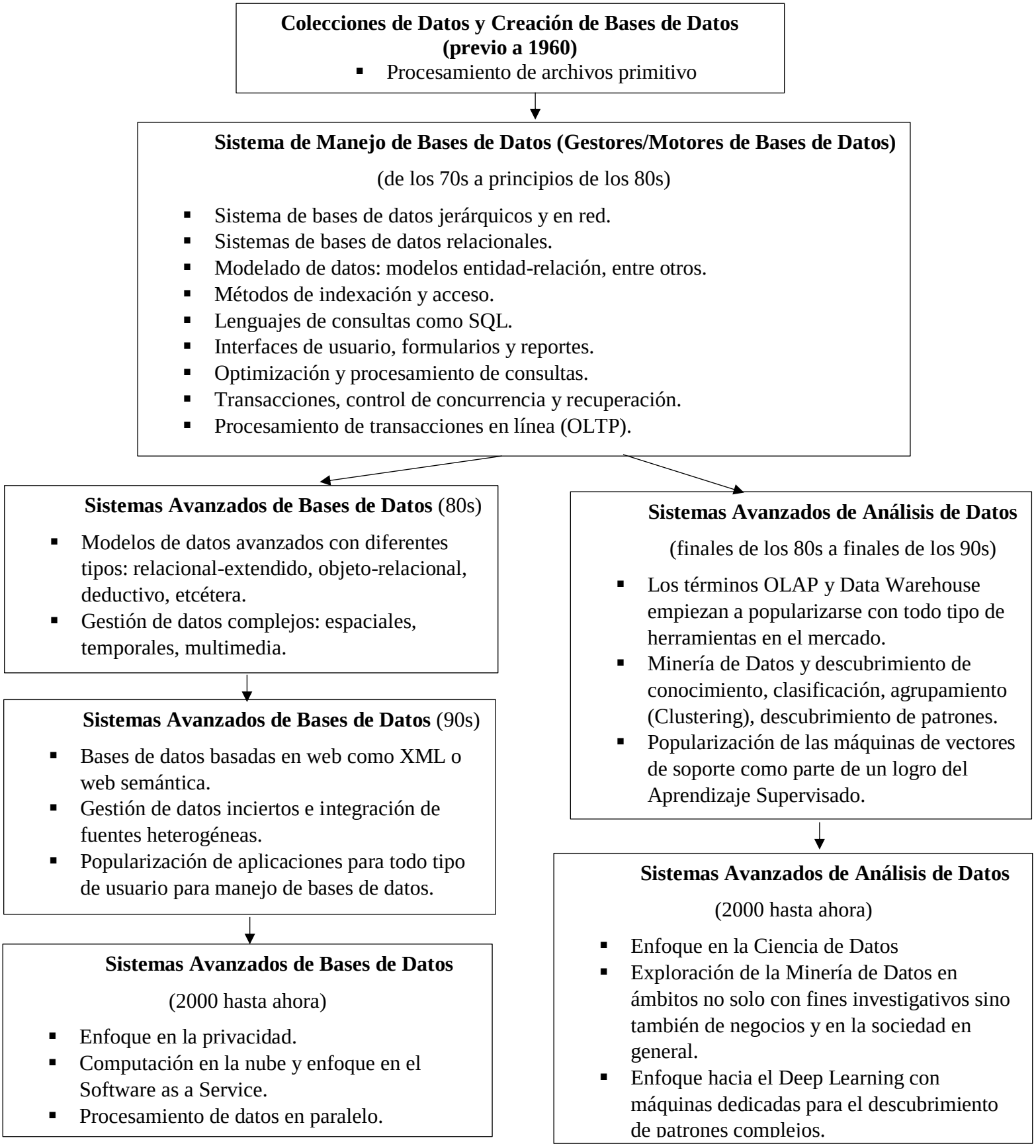


Figura 3. Historia de las colecciones de Datos. En Data Mining: Concepts and Techniques (p. 3), por Han, Kamber y Pei.

1.3 Ciencias y Campos Relacionados con la Minería de Datos

La principal ayuda de la Minería de Datos es la Estadística, que es una rama de las Matemáticas que busca la organización, recolección y análisis de datos. Como menciona Tufféry (2011), la Minería de Datos y la Estadística al juntarse en un inicio tuvieron un campo específico que constaba de ayuda para investigación de laboratorio, ensayos clínicos, estudios actuariales y análisis de riesgos; es decir, a pesar de que su valor nunca estuvo negado, se restringía su aplicación por temas de desarrollo de la industria y disponibilidad de equipos, tal y como se comentó en la sección anterior. En la actualidad, el campo de aplicación se encuentra en constante expansión que va desde lo infinitamente pequeño (Genómica) hasta lo infinitamente grande (Astrofísica); desde lo más general (gestión de la relación con el cliente) hasta lo más específico (asistencia a pilotos durante un vuelo); desde lo más público (comercio electrónico) hasta lo más privado (prevención del terrorismo, detección de fraudes en telefonía móvil, robos en tarjetas bancarias); desde lo más teórico (ciencias humanas, biología, medicina y farmacología) hasta lo más práctico (control de calidad, gestión de producción, etcétera); pasando por predicciones de audiencia para un programa de televisión, o controles en temas alimentarios.

Simplemente con la mención de las aplicaciones de la Minería de Datos y Estadística se puede notar que cubren un gran espectro en las necesidades de la sociedad actual. La Minería de Datos usa gran parte de las técnicas que provee la Estadística y las aplica mediante la Informática (vía la construcción de software para automatizar procesos y gestionar de mejor forma los datos y la información encontrada). En conjunto, estas dos ramas de la Ciencia se encargan de analizar grandes volúmenes de datos, a veces con el objetivo de una toma de decisiones rápida, como en el caso de algunos de los ejemplos dados anteriormente. Se puede decir que, la asistencia en la toma de decisiones se está convirtiendo en un objetivo primario de la Minería de Datos y la Estadística. Este enfoque de apuntar a la toma de decisiones no es algo nuevo; de hecho, se lo viene usando ampliamente en la Medicina en donde la Minería de Datos permite limitar la subjetividad humana cuando se trata de algún caso clínico que no tenga muchos precedentes, como alguna variante de un virus o alguna enfermedad poco conocida.

Sin embargo, en el mundo general que vivimos hoy, la riqueza de un negocio se encuentra en sus clientes, la cuota de cliente ha sustituido la cuota de mercado. Las empresas líderes han sido valoradas en términos de sus archivos de clientes, sobre la base de que cada

cliente vale mucho dinero. En este contexto, comprender las expectativas de los clientes y anticipar sus necesidades se convierte en un objetivo importante de muchas empresas que desean aumentar su rentabilidad y la lealtad de sus clientes, mientras controlan el riesgo y usan los canales adecuados para vender un producto correcto en un momento correcto.

En el ámbito comercial existe una serie de preguntas que se deben plantear, más allá de las clásicas: "¿cuántos clientes han comprado este producto en este período?" Sino también "¿cuál es su perfil?", "¿en qué otros productos están interesados?" Y "¿cuándo se interesan?". Los perfiles que se tienen que descubrir son generalmente complejos, no se trata solamente de dar con temas como: mayor-menor, hombre - mujer, urbano - rural, es decir, temas que se pueden "deducir" haciendo uso a las estadísticas descriptivas; sino categorías con combinaciones más complicadas, en las que se tienen variables discriminantes que no son necesariamente lo se pudo haber predicho en una etapa inicial del análisis, y por ende no se pudieron encontrar por casualidad; esto es especialmente claro en el caso de comportamientos "raros" o fenómenos. Con la minería de datos, se puede pasar del análisis confirmatorio al análisis exploratorio. (Tufféry, 2011).

Los métodos de minería de datos son ciertamente más complejos que los de la Estadística Descriptiva. Se basan en herramientas de inteligencia artificial (redes neuronales), teoría de la información (árboles de decisión), teoría del aprendizaje automático y, sobre todo, estadísticas inferenciales y análisis de datos "convencionales", incluido el análisis factorial, la agrupación y el discriminante. análisis, etc. (Tufféry, 2011)

Las razones por las que la minería de datos se ha trasladado de las universidades y los laboratorios de investigación al mundo de los negocios incluyen presiones de la competencia y las nuevas expectativas de los consumidores, así como los requisitos reglamentarios en algunos casos, como los productos farmacéuticos (donde los medicamentos deben probarse antes de ser comercializados, o los bancos (donde el patrimonio debe ajustarse de acuerdo con la cantidad de exposición y el nivel de riesgo incurrido). Este desarrollo ha sido posible gracias a tres grandes avances técnicos.

Evidentemente, como se ha comentado, la Minería de Datos se relaciona de una u otra manera con las tecnologías informáticas, debido al desarrollo de software para ejecutar mediante estos programas los métodos de Minería de Datos previamente ideados. Un especialista en TI verá un modelo de minería de datos como una aplicación de TI, en otras palabras, un conjunto de

instrucciones escritas en un lenguaje de programación para llevar a cabo ciertos procesos, con lo siguiente:

- Proporcionar datos de salida que resumen los datos de entrada.
- Proporcionar datos de salida de un nuevo tipo, deducidos de los datos de entrada y utilizados para la toma de decisiones.

Al igual que todas las aplicaciones de TI, una aplicación de minería de datos pasa por varias fases:

- Desarrollo o construcción del modelo en el entorno de toma de decisiones.
- Pruebas o verificación del rendimiento del modelo en el entorno de toma de decisiones.
- Puesta en producción o aplicación del modelo a los datos de producción para obtener los datos de salida especificados. (Tufféry, 2011)

Sin embargo, no se puede tratar a un desarrollo de Minería de Datos como si se tratara de cualquier tipo de aplicación informática; estos desarrollos tienen algunas características distintivas, como las siguientes:

- La fase de desarrollo no se puede completar en ausencia de datos, en contraste con un desarrollo de TI que tiene lugar de acuerdo con una especificación o requerimiento; el desarrollo de un modelo depende principalmente de los datos (incluso si también hay una especificación).
- El desarrollo y las pruebas se llevan a cabo en el mismo entorno, con solo los conjuntos de datos que difieren entre sí.
- Para obtener un modelo óptimo, es normal y necesario moverse frecuentemente entre las pruebas y el desarrollo; Algunos programas controlan estos movimientos de manera en gran parte automática para evitar cualquier pérdida de tiempo.
- El análisis de datos para el desarrollo y las pruebas se realiza mediante un programa de propósito especial, generalmente diseñado por SAS, SPSS que forma parte IBM, KXEN, Statistica o SPAD, o software de código abierto. (Tufféry, 2011).

1.4 Herramientas FLOSS o FOSS

FOSS significa Software libre y de código abierto (free and open source software) y es un término general para software que se considera simultáneamente libre y de código abierto. Estos software permiten al usuario inspeccionar el código fuente y proporcionan un alto nivel de control de las funciones del programa, en comparación con el software propietario. Usualmente la gente confunde el término "software libre" con "software gratis", es decir, no se refiere al costo monetario sino a la licencia que mantiene con las libertades civiles del usuario del software.

La elección de las herramientas FLOSS para este trabajo no es algo elegido al azar. El rápido crecimiento del uso de herramientas open source, por ejemplo, pone en evidencia muchas de las ventajas que se puede obtener al trabajar con este modelo de negocio. Como bien lo menciona Coppola & Neelley (2004), hay ciertas preeminencias cuando se trabaja con software open source:

- Se presenta una evolución constante y orgánica del código, debido a la gran cantidad de individuos con acceso, trabajando continuamente en actualizar el o los sistemas mediante una serie de desarrollos en paralelo y que terminan pasando por la revisión de un equipo central, además de las obvias y gratuitas colaboraciones que realizan internautas a través de foros en línea.
- El mismo punto anterior deriva en una gestión más eficaz para cubrir posibles necesidades de los clientes; una persona reporta un determinado error o malfuncionamiento, éste es corregido por alguno de los equipos de desarrolladores que trabajan en el código fuente, y el software finalmente se puede mejorar, parcheando los defectos antes mencionados.
- Los problemas de seguridad pueden ser un problema menor al tener tantos involucrados, con el código fuente expuesto a todo el mundo. Debido a esto, los mecanismos de seguridad tienden a ser más estrictos pues cualquier persona podría analizar el código y descubrir maneras de sobrepasar las barreras impuestas.

Coppola & Neelley (2004) sugieren que muchas de las personas que aseveran estar en contra de este tipo de desarrollos, donde el código fuente está a ojos de todo el mundo, encuentran respaldo en que es una mala idea con respecto a los negocios, puesto que muchas personas podrían

hacer uso del código para el desarrollo de más y mejores productos de software; aún más si hablamos de software gratuito. El punto es que, evidentemente existen maneras de hacer negocio con desarrollos FLOSS:

- Establecer un modelo de negocios basado en publicidad, como la mayoría de las apps para teléfonos inteligentes. Este punto puede que no sea del agrado de todos los consumidores, pero sí que resulta una alternativa para ofrecer software de código abierto y gratuito.
- Cuando hablamos de sistemas más personalizados, se puede hacer negocio en servicios posteriores a la codificación, como cobrar por la implementación del software en un determinado entorno y, sobre todo, basar el modelo de negocios en el mantenimiento.
- Establecer un modelo de negocios basado en la capacitación de los usuarios en un determinado software. Puede que el negocio de un desarrollo FLOSS se encuentre en el cobro por enseñar a usar el sistema, ofrecer entrenamiento a los distintos usuarios que así lo requieran.

2. Técnicas de Minería de Datos

2.1 Descripción del Proceso General de un proyecto de Minería de Datos

Como se había especificado en el capítulo anterior, el proceso de minería de datos es iterativo y se consta de algunos pasos, o más específicamente, fases. De acuerdo con Olson & Delen (2008), con el objetivo de llevar a cabo sistemáticamente el análisis de minería de datos, generalmente se cumple con el proceso general; se han generado dos estándares:

- CRISP-DM (Cross-Industry Standard Process for Data Mining), que es un proceso estándar de la industria y consiste en una secuencia de pasos que generalmente se usan en un estudio de Minería de Datos
- SEMMA, que está orientado a SAS (Analytics, Bussiness Intelligence y Data Management)

Los pasos que se describen a continuación están orientados a CRISP-DM.

2.1.1 Definición del Problema

Un proyecto de Minería de Datos arranca entendiendo el problema del negocio. No necesariamente se debe generalizar estos proyectos para casos empresariales, pero la ejecución de estas fases se explicará de esta manera para simplificar los temas.

Los expertos en Minería de Datos, los expertos del negocio y todos los demás relacionados se juntan y laboran en estrecha colaboración para definir los objetivos del proyecto y los requisitos desde una perspectiva empresarial. El objetivo del proyecto se traduce luego en una definición de problema de minería de datos. En la fase de definición del problema, las herramientas de minería de datos aún no son necesarias.

El elemento clave de un estudio de Minería de Datos es saber para qué sirve el estudio. Esto comienza con una necesidad gerencial de nuevos conocimientos y una expresión del objetivo de negocios con respecto al estudio que se realizará.

Como se dijo anteriormente, estos pasos se los enfoca en temas de negocios, pero un proyecto de Minería de Datos no necesariamente está ligado al sector empresarial, puede resultar en un trabajo netamente investigativo y sin fines de lucro.

Cuando se empieza un proyecto, se necesitan resolver algunas cuestiones como: “¿qué tipos de clientes están interesados en cada uno de nuestros productos?”, “¿cuáles son los perfiles típicos de nuevos clientes y cuánto valor nos proporcionan?”. Luego se debe desarrollar un plan

para encontrar dicho conocimiento, recopilando datos, analizando dichos datos y emitiendo informes. (Olson & Delen, 2008)

Evidentemente cuando se habla de negocios se tiene que establecer un presupuesto, para que las compañías determinen la factibilidad del estudio. En ambientes más pequeños, con estudios sin ánimo de beneficio económico, el presupuesto es un tema que puede obviarse.

2.1.2 Exploración de los datos

La exploración de los datos no se trata de entender externamente la situación que los mismos conllevan, sino del trabajo de expertos en metadatos que conocen el arte de recopilar, describir y explorar datos, así como la identificación de problemas de calidad de los datos. En este punto, las herramientas tradicionales de análisis de datos, como la Estadística, se usan para explorar los datos.

Una vez que se establecen los objetivos iniciales, así como el plan del proyecto, la comprensión de los datos considera los requisitos de los datos. Esta fase puede incluir una recopilación inicial de los datos, su descripción, exploración y algo muy importante, la verificación de la calidad de los datos. La exploración como la visualización de estadísticas de resumen (que incluye la visualización de variables categóricas) puede ocurrir cuando se concluye esta fase. Los modelos como el análisis de clúster también se pueden aplicar en este paso con la intención de identificar patrones en los datos o encontrar predicciones de los mismos.

Se puede decir que, en esta etapa del proceso se deben seleccionar datos relacionados de muchas fuentes (como bases de datos) disponibles para describir correctamente el problema en específico.

Sin embargo, existen algunos contratiempos como la realización de una descripción lo suficientemente clara y concisa del eje central del proyecto. Por ejemplo, mediante Minería de Datos se busca identificar patrones de bancarrota de los titulares de tarjetas de crédito. Otros problemas como la correcta identificación de datos, o el aseguramiento de independencia de las variables elegidas, deberán ser solucionadas a su debido tiempo.

Las fuentes de datos para la selección de datos pueden variar. Normalmente, los tipos de fuentes de datos para aplicaciones comerciales incluyen datos demográficos (como ingresos, educación, número de hogares y edad), datos socio gráficos (como pasatiempo, membresía de clubes y entretenimiento), datos transaccionales (registros de ventas, crédito). Gastos de tarjeta, cheques emitidos), etc. El tipo de datos puede categorizarse como datos cuantitativos y

cualitativos. Los datos cuantitativos son medibles utilizando valores numéricos. Puede ser discreto (como los enteros) o continuo (como los números reales). Los datos cualitativos, también conocidos como datos categóricos, contienen datos nominales y ordinales. Los datos nominales tienen valores finitos no ordenados, como los datos de género que tienen dos valores: masculino y femenino. Los datos ordinales tienen valores ordenados finitos. Por ejemplo, las calificaciones crediticias de los clientes se consideran datos ordinales, ya que las calificaciones pueden ser excelentes, justas y malas. (Olson & Delen, 2008).

Los datos cuantitativos se pueden representar fácilmente mediante algún tipo de distribución de probabilidad. Una distribución de probabilidad describe cómo se dispersan y se forman los datos. Por ejemplo, los datos normalmente distribuidos son simétricos, y comúnmente se les conoce como forma de campana. Los datos cualitativos pueden codificarse primero en números y luego describirse mediante distribuciones de frecuencia. Una vez que los datos relevantes se seleccionan de acuerdo con el objetivo comercial de la minería de datos, se debe continuar el procesamiento previo de los datos. (Olson & Delen, 2008)

2.1.3 Preparación de datos

El propósito del preprocesamiento o preparación de datos es tratar de limpiar los datos seleccionados para una mejor calidad. Algunos datos seleccionados pueden tener diferentes formatos porque se eligen de diferentes fuentes de datos. Si los datos seleccionados provienen de archivos planos, mensajes de voz o texto web, deben convertirse a un formato que sea consistente. Simplificando, la limpieza de datos significa: filtrar, agregar y completar los valores faltantes (proceso conocido como imputación). Al filtrar los datos, los datos seleccionados se examinan en busca de valores atípicos y redundancias. (Olson & Delen, 2008)

Por ejemplo, si el ingreso de un cliente incluido en la clase media se establece como $ 1000000, es evidente que se trata de un error y debe sacarse del proyecto de minería de datos que examina los diversos aspectos de la clase media, ya que podría causar problemas en futuros análisis. Los valores atípicos pueden ser causados por muchas razones, como errores humanos o errores técnicos, o pueden ocurrir naturalmente en un conjunto de datos debido a eventos extremos. Supongamos que la edad del titular de una tarjeta de crédito se registra como "12". Es probable que se trate de un error humano, puesto que en la mayoría de las circunstancias no existen personas menores de edad con tarjetas de crédito. Sin embargo, en realidad podría haber un preadolescente adinerado e independiente con importantes hábitos de compra. La eliminación

arbitraria de este valor atípico podría descartar información valiosa. Los datos redundantes son la misma información registrada de varias maneras diferentes. Las ventas diarias de un producto en particular son redundantes a las ventas estacionales del mismo producto, porque podemos derivar las ventas de datos diarios o estacionales. El objetivo de todo esto es encontrar un proceso de análisis práctico que permita identificar correctamente los datos que deben ser o no eliminados. (Olson & Delen, 2008).

Al agregar datos, las dimensiones de los datos se reducen para obtener información agregada. Tenga en cuenta que, aunque un conjunto de datos agregados tiene un volumen pequeño, la información permanecerá. Si se considera una promoción de marketing para la venta de muebles en los próximos 3 o 4 años, entonces los datos de ventas diarias disponibles se pueden agregar como datos de ventas anuales. El tamaño de los datos de ventas se reduce dramáticamente. Al suavizar los datos, se encuentran los valores faltantes de los datos seleccionados y luego se agregan valores nuevos o razonables. Estos valores agregados podrían ser el número promedio de la variable (media) o el modo. Un valor faltante a menudo no causa ninguna solución cuando se aplica un algoritmo de extracción de datos para descubrir los patrones de conocimiento.

Los datos se pueden expresar en varias formas diferentes, como valores numéricos, texto, tipos distintos de datos, binarios, etcétera; con esto se quiere hacer notar que no necesariamente los datos son números, pueden muchas veces ser binarios (como si una persona está casada o no).

2.1.4 Modelado

Empleando ciertas técnicas, así como herramientas de Minería de Datos se pueden desarrollar modelos que permitan obtener una relación entre los datos como: patrones, predicciones, asociaciones entre variables, etcétera.

Las herramientas como la inducción de reglas generalizadas pueden desarrollar reglas de asociación iniciales. Una vez que se obtiene una mayor comprensión de los datos (a menudo a través del reconocimiento de patrones activado al ver la salida del modelo, o también mediante la obtención de predicciones.), se pueden aplicar modelos más detallados que sean adecuados al tipo de datos. La división de datos en entrenamiento y pruebas es también necesaria para el modelado. (Olson & Delen, 2008).

En otras palabras, es aquí cuando los expertos en Minería de Datos seleccionan y aplican varias funciones de minería (debido a que se pueden usar diferentes funciones para el mismo tipo

de problema). Algunas de las funciones requieren datos específicos, por lo que los expertos deben evaluar cada modelo y usar el que consideren más propicio.

La fase de modelado y la fase de evaluación están acopladas. Se pueden repetir varias veces para cambiar los parámetros hasta lograr los valores óptimos. Cuando se completa la fase final de modelado, se presupone que se ha construido un modelo de alta calidad.

La Minería de Datos puede ser completada por medio de varias técnicas como la Regresión Lineal, Asociación, Clasificación, Clustering, Predicciones, Patrones Secuenciales, etcétera. Varias de estas técnicas serán revisadas a profundidad posteriormente en este capítulo.

2.1.5 Evaluación

Los resultados del modelo deben evaluarse en el contexto de los objetivos comerciales establecidos en la primera fase (comprensión). Esto conducirá a la identificación de otras necesidades (a menudo a través del reconocimiento de patrones o evaluación de predicciones), volviendo con frecuencia a anteriores fases de CRISP-DM. Adquirir comprensión comercial es un procedimiento iterativo en la minería de datos, donde los resultados de varias herramientas de visualización, estadísticas e inteligencia artificial muestran al usuario nuevas relaciones que proporcionan una comprensión más profunda de las operaciones organizacionales. (Olson & Delen, 2008).

Los expertos deben evaluar el modelo generado; en el caso de que éste no satisfaga sus expectativas, se debe regresar a la fase de modelado y reconstruir el modelo cambiando sus parámetros hasta lograr valores óptimos.

Cuando finalmente estén satisfechos con el modelo, pueden extraer explicaciones de negocios y evaluar las siguientes preguntas:

¿El modelo logra el objetivo planteado inicialmente?

¿Se han considerado todas las variables alrededor del problema?

2.1.6 Lanzamiento

La Minería de Datos se puede utilizar tanto para verificar hipótesis mantenidas anteriormente como para el descubrimiento de conocimiento (identificación de relaciones inesperadas y útiles). A través del conocimiento descubierto en las fases anteriores del proceso antes mencionado, se pueden obtener modelos del tipo CRISP-DM 11 que luego se pueden aplicar a las operaciones comerciales para muchos propósitos, incluida la predicción o

identificación de situaciones clave. Estos modelos deben ser monitoreados para detectar cambios en las condiciones de operación ya que lo que podría ser cierto hoy, puede no serlo dentro de un año. (Olson & Delen, 2008).

Si ocurren cambios significativos, el modelo debe ser rehecho. También es aconsejable registrar los resultados de los proyectos de Minería de Datos para que haya evidencia documentada disponible para estudios futuros.

2.2 Aprendizaje Supervisado y No supervisado

2.2.1 Aprendizaje Supervisado

El Aprendizaje Supervisado es un tipo de aprendizaje automático que se basa en que una máquina pueda aprender una manera de convertir determinada entrada en una salida basada en un conjunto de datos de ejemplo.

Tal y como mencionan Suresh, Sundararajan , & Savitha (2012), y hablando desde su enfoque hacia las Redes Neuronales Complejas, el aprendizaje dentro de una red neuronal se puede llamar "supervisado" siempre que ocurra con un maestro; éste maestro tiene conocimiento del entorno, es decir, posee una serie de entradas y salidas que funcionan como ejemplos y que se suelen denominar "dataset de entrenamiento".

Según Russell & Norvig (2010), en el aprendizaje supervisado, un agente (que no es más que una entidad que recibe percepciones de su entorno y en base a ella realiza ciertas acciones) observa algunos pares de entrada-salida como ejemplos y aprende una función que asigna una salida en base a una entrada específica. Un ejemplo práctico para comprender cómo funciona el Aprendizaje Supervisado sería el siguiente: imaginemos que el agente es un estudiante que desea aprender a conducir un vehículo. Éste estudiante posee un maestro que le guía durante su aprendizaje; las entradas, percepciones y la salida son provistas por el maestro, que dice a su alumno "¡Frena!" o "¡Vira a la izquierda!". También podríamos pensar en un ejemplo similar donde se tengan otros elementos como una cámara dentro del vehículo que provee de imágenes importantes para el estudiante y que sirven como entradas y la salida es el mismo maestro diciendo algo similar a "¡Eso es un autobús!". El problema surge cuando, en la vida real, las distinciones no son tan nítidas y no se puede distinguir correctamente entre los datasets de entrenamiento. Por ejemplo, en el aprendizaje semi-supervisado se nos dan algunos ejemplos etiquetados y se debe hacer lo que se pueda con un gran conjunto de ejemplos no etiquetados.

Inclusos las propias etiquetas pueden ser falsas o distintas a lo que se esperaría. Imaginando otro ejemplo, donde alguien desee recopilar datos de un grupo de personas en base a una fotografía y preguntándoles directamente su edad, pero en realidad algunas personas mintieron a la hora de decir su edad. No es solamente que se tenga que lidiar con un ruido aleatorio, sino más bien con inexactitudes sistemáticas que no son posibles de resolver con aprendizaje supervisado ya que involucra el análisis de imágenes, edades auto informadas y edades reales que son desconocidas. Este pequeño problema nos sugiere que la línea que separa el Aprendizaje Supervisado del No Supervisado realmente es muy pequeña y en ciertos casos debe cruzarse para poder cumplir con el objetivo de un determinado problema.

El objetivo del aprendizaje supervisado se puede resumir de la siguiente manera, haciendo una aproximación matemática:

Dado un conjunto de entrenamiento de "N" pares de ejemplo entrada-salida de la forma:

$$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$$

Donde cada y_j fue generado por una función desconocida de la forma $y = f(x)$, encontrar una función similar "h" que se aproxime lo más posible a la función original "f".

En este caso cabe aclarar que x y y pueden tomar valor, no necesariamente numérico. La función "h" vendría a representar una especie de hipótesis. Entonces, podemos decir que el "aprendizaje" es una búsqueda en el espacio de las posibles hipótesis hasta dar con una que se desempeñe de la manera esperada, incluso con ejemplos nuevos que no estén incluidos en el conjunto de entrenamiento. Para medir la efectividad de la nueva función "h", es decir, para medir la precisión de una determinada hipótesis, se manejan sets o conjuntos de datos de prueba que sean distintos del conjunto de entrenamiento original. Se dice que una hipótesis generaliza bien si predice correctamente el valor de "y" para ejemplos nuevos. A veces, la función "f" es estocástica, es decir, no es estrictamente una función de x, por lo que se tiene que encontrar una distribución de probabilidad condicional P(Y|x). (Russell & Norvig, 2010)

Cuando la salida y es uno de los valores de un conjunto finito de datos (como por ejemplo el conjunto: ["soleado", "nublado", "lluvioso"]), el problema de aprendizaje es conocido como "Clasificación", y se tiene una subdivisión llamada "Clasificación binaria o booleana" en el caso de que el conjunto tenga solo dos valores. (Russell & Norvig, 2010)

Se profundizará en los conceptos de "Regresión" y "Clasificación" más adelante en este capítulo.

Cuando la salida y es un número (como por ejemplo el valor de la temperatura del día de mañana), el problema de aprendizaje es llamado "Regresión"; al resolver un problema de regresión lo que se busca es una expectativa condicional o un valor promedio de y porque la probabilidad de que hayamos encontrado exactamente el número correcto de valores reales para y es nula o igual a cero. (Russell & Norvig, 2010)

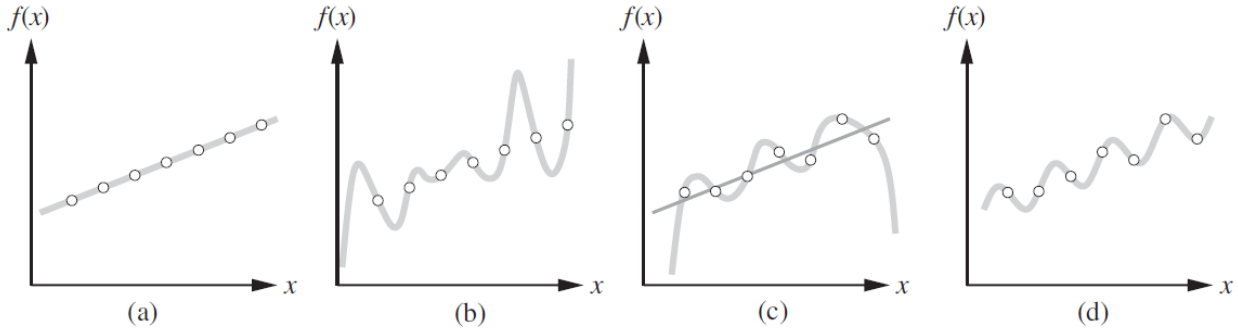


Figura 4. Ejemplos de pares x,y con sus respectivas hipótesis. En Artificial Intelligence: A Modern Approach (p. 696), por Russell y Norvig.

En la figura 4 se muestra un ejemplo donde se busca ajustar una función de una sola variable a algunos puntos de datos. Los ejemplos están en el plano (x, y), donde $y = f(x)$. No se sabe qué función exactamente es, pero se puede aproximar dicha función con otra similar "h" que fue seleccionada del "espacio de hipótesis" H, que en este ejemplo resulta ser un conjunto de polinomios, como $x^5 + 3x^2 + 2$. En el ejemplo a) se tiene un ajuste exacto por una línea recta con el polinomio $0.4x + 3$; esta línea es llamada "hipótesis consistente" porque concuerda con todos los datos. En el ejemplo b) se tiene un polinomio de alto grado que también es consistente con los datos. Esta consistencia con dos ejemplos nos lleva a un problema, ¿cómo elegir entre múltiples hipótesis consistentes? La respuesta está en el principio llamado "La Navaja de Ockham" del filósofo inglés William Ockham y no es otra que la simpleza, es decir, cuando se tengan múltiples hipótesis consistentes, se debe elegir la más simple de todas. El problema derivado de esto es que definir "simpleza" no es para nada sencillo, sin embargo, en este caso específico se puede ver con claridad que un polinomio de primer grado es mucho más simple que un polinomio de grado 7, por lo que (en este caso en particular) se debería optar por la opción a. (Russell & Norvig, 2010).

Para el caso c) se muestra un segundo conjunto de datos. No existe una línea recta consistente para estos datos y, de hecho, se requiere de un polinomio de grado 6 para el ajuste exacto. Como solo existen 7 puntos de datos, un polinomio con 7 parámetros no parece encontrar

ningún patrón en los datos y así no se puede esperar que generalice bien. Una línea recta que no es coherente con ninguno de los puntos de datos pero que puede generalizar bastante bien para valores desconocidos de x también se muestra en el ejemplo c). (Russell, Norvig ,2010)

En cuanto al caso d) el espacio de hipótesis es ampliado para admitir polinomios tanto de x como de sen(x); haciendo este pequeño ajuste al espacio de hipótesis, se encuentra que la solución de c) era más simple de lo esperado, siendo la función: $ax + b + csen(x)$. El análisis se enfoca a que, la elección del espacio de hipótesis es muy importante pues puede eliminar posibles resultados que se ajusten perfectamente a la función original y esto, como vimos previamente, complica mucho el trabajo. Se dice que un problema de aprendizaje es "realizable" si el espacio de hipótesis contiene al menos una función verdadera, aunque desafortunadamente, no siempre se puede decir si cierto problema de aprendizaje es realizable porque las funciones verdaderas no se conocen. (Russell & Norvig, 2010).

2.2.2 Aprendizaje no Supervisado

2.2.2.1 Definición

Como lo menciona Ghahramani (2004), consideremos una máquina (o un ser viviente) que recibe algunas secuencias de entradas $x_1, x_2, x_3, \dots, x_t$ donde x_t es la entrada en el tiempo *t*. Esta entrada, comúnmente llamada "dato", puede corresponder a una imagen obtenida por los ojos de un ser, los píxeles de una cámara, etcétera. Pero también puede corresponder a datos sensoriales menos obvios de identificar como las palabras en una noticia, la lista de productos por comprar en un supermercado, etcétera. Tal y como se pudo analizar en la sección previa, en el Aprendizaje Supervisado la máquina además de las entradas que recibe, también cuenta con una secuencia de salidas deseadas del tipo y_1, y_2, ...; por lo que el objetivo de la máquina no es más que el de aprender a generar una salida correcta dada una nueva entrada; siendo que ésta salida puede ser una etiqueta de clase (en Clasificación) o un número real (en Regresión).

En cuanto al Aprendizaje no Supervisado, la máquina simplemente recibe las entradas x_1, x_2, x_3, ..., pero sin ningún conjunto de salidas deseadas. Presentadas las condiciones, puede parecer un tanto misterioso imaginar qué es lo que la máquina puede aprender dado que no recibe retroalimentación alguna por parte de su entorno. Sin embargo, sí es posible desarrollar un marco formal para el Aprendizaje no Supervisado basándonos en la noción de que el objetivo de la máquina es construir representaciones de la entrada que se pueden utilizar para la toma de

decisiones, así como predecir entradas futuras o comunicar de manera eficiente las entradas dirigidas a otra máquina, etcétera. En cierto modo, podemos decir que este tipo de aprendizaje de máquina puede considerarse como la tarea de encontrar patrones en los datos o predecir nuevos valores, que van más allá de lo que se consideraría ruido puro sin estructurar. Dos ejemplos clásicos y simples de Aprendizaje no Supervisado son el Clustering y la reducción dimensional. (Ghahramani, 2004).

2.2.2.2 Relación con la Estadística y la Teoría de la Información

Casi todo el trabajo relacionado con Aprendizaje no Supervisado puede entenderse como el aprender un cierto modelo probabilístico de los datos. Incluso si a la máquina no se le ha dado ningún conjunto de salidas deseadas, puede tener sentido para la máquina estimar un modelo que representen la distribución de probabilidad para una nueva entrada x_t dado un conjunto de entradas del tipo $x_1, x_2, x_3, \dots, x_{t-1}$. Con lo que, se obtiene un modelo del tipo $P\ (x_t|x_1, \dots, x_{t-1})$. En casos más simples donde el orden en que los datos de entrada llegan es irrelevante o desconocido, entonces la máquina puede crear un modelo en donde asume que los datos x_1, x_2, ... son extraídos de forma independiente e idéntica de alguna distribución de la forma $P(x)^2$. (Ghahramani, 2004)

El modelo anteriormente mencionado se podría utilizar para la detección o supervisión de valores atípicos. Supongamos que x representa patrones de lectores de sensores de alguna planta nuclear y asumamos que P(x) aprende de los datos recopilados de alguna planta nuclear que sí funciona normalmente. Este modelo se puede usar para evaluar la probabilidad de una nueva lectura del sensor; por lo que, asumimos que si la probabilidad es más baja que lo que se consideraría "normal" y asumiendo que se ha probado que el modelo es eficiente, entonces se puede llegar a la conclusión de que la planta nuclear no está comportándose de la manera esperada. Un modelo probabilístico también se puede usar para la Clasificación. Suponiendo que P1(x) es un modelo que cuenta con los atributos de los titulares de tarjetas de crédito que pagaron su cuota a tiempo, y P2(x) es de aquellos que incumplieron sus pagos, al evaluar a una futura nueva solicitante, la máquina podrá decidir clasificarla en una de estas dos categorías. (Ghahramani, 2004)

Con un modelo probabilístico también se puede lograr una comunicación eficiente y una mejor comprensión de los datos. Imaginemos que queremos transmitir, a través de una línea de comunicación digital, símbolos x extraídos al azar de P(x). Por ejemplo, x puede ser una letra del

alfabeto, o una imagen, y la línea de comunicación puede ser Internet. Intuitivamente debemos codificar nuestros datos para que los símbolos que aparecen con mayor frecuencia tengan palabras clave con menos bits, de lo contrario, estamos desperdiciando el ancho de banda. El Primer Teorema de Shannon lo cuantifica diciéndonos que el número óptimo de bits que se debe usar para codificar un símbolo con probabilidad P(x) es $-log_2 P(x)$. Usando este número de bits para cada símbolo, el costo de codificación esperado resulta ser la entropía de la distribución P.

$$H(P) \stackrel{\text{def}}{=} -\sum_X P(x) log_2 P(x)$$

}En general, la verdadera distribución de los datos es desconocida, pero se puede plantear un modelo de esta distribución; llamaremos a este posible modelo como Q(x). El código óptimo con respecto a este modelo usaría bits del tipo $-log_2 Q(x)$ y el costo de codificación esperado (teniendo expectativas con respecto a la distribución real), sería: $-\sum_x P(x) log_2 Q(x)$. Por lo tanto, ya tenemos dos costos de codificación (uno con respecto a la distribución original y otro tomando en cuenta algún modelo que se acerque a dicha distribución). La diferencia entre estos dos costos de codificación es llamada la Divergencia Kullback-Leibler (KL) y se la denota de la siguiente manera:

$$KL(P||Q) \stackrel{\text{def}}{=} \sum_x P(x) log \frac{P(x)}{Q(x)}$$

La Divergencias KL es positiva y cero sí y solo sí P=Q. Esto mide la falta de ineficiencia de codificación en bits al utilizar un modelo Q para comprimir datos cuando la verdadera distribución de datos es P. Por lo tanto, cuanto mejor sea el modelo de datos, más eficientemente se podrá comprimir y comunicar nuevos datos. Este es uno de los vínculos más importantes entre el Aprendizaje Automático, la Estadística y la Teoría de la Información. (Ghahramani, 2004)

2.2.2.2.1 El Teorema de Bayes

Dentro de la Estadística, encontramos un tema especialmente útil para el Aprendizaje Automático, que es el Teorema de Bayes o la Regla de Bayes que dicta lo siguiente:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Esta regla se desprende de la igualdad $P(x,y) = P(x)P(y|x) = P(y)P(x|y)$ y se puede usar para motivar un marco estadístico coherente para el Aprendizaje Automático. La idea básica es la siguiente: imaginemos que se desea construir una máquina que tengan creencias sobre el

mundo y que actualice estas creencias sobre la base de los datos observados. La máquina debe representar numéricamente las fortalezas de sus creencias. Se he demostrado que, si acepta ciertos axiomas de inferencias coherentes, conocimos como "Los Axiomas de Cox", entonces se obtiene un resultado notable: si la máquina debe representar la fuerza de sus creencias con números reales, entonces la única forma razonable y coherente de manipular dichos pensamientos es hacer que satisfagan las reglas de probabilidad, como el Teorema de Bayes. Por lo tanto, $P(X = x)$ se puede usar no solo para representar la frecuencia con la que la variable X toma el valor "x", sino también para representar el grado de creencia de que X = x. (Ghahramani, 2004)

De la Regla de Bayes podemos derivar el siguiente marco de trabajo simple para Aprendizaje Automático. Asumamos un universo de modelos Ω, siendo $\Omega = \{1, \dots, M\}$ aunque no debe ser un conjunto finito o incluso contable. Las máquinas empiezan con algunas creencias sobre modelos $m \in \Omega$ como, por ejemplo: $\sum_{m=1}^{M} P(m) = 1$. Un modelo es simplemente una distribución de probabilidad sobre ciertos datos, en otras palabras, $P(x|m)$. Para simplificar la explicación, supongamos que en todos los modelos los datos que se toman son independientes e idénticamente distribuidos. Después de observar el dataset $D = \{x_1, \dots, x_N\}$ las creencias acerca de los modelos están dadas por:

$$P(m|D) = \frac{P(m)P(D|m)}{P(D)} \propto P(m)\prod_{n=1}^{N} P(x_n|m)$$

Lo que leemos como modelos posteriores es el anterior multiplicado por la probabilidad, normalizado. La distribución predictiva sobre nuevos datos, que se utilizarían para codificar nuevos datos de manera eficiente, es: $P(x|D) = \sum_{m=1}^{M} P(x|m)P(m|D)$

De nuevo, esto se deduce de las reglas de la Teoría de la Probabilidad y del hecho de que se supone que los modelos solo producen datos que son independientes e idénticamente distribuidos. A menudo, los modelos se definen escribiendo una distribución de probabilidad paramétrica, por lo tanto, el modelo m podría tener parámetros θ, que se supone son desconocidos (aunque bien podría tratarse de un conjunto o vector de parámetros). Para que un modelo se considere como "bien definido" desde la perspectiva del aprendizaje bayesiano, uno tiene que definir una prioridad sobre estos parámetros del modelo $P(\theta|m)$ que naturalmente tiene que satisfacer la siguiente igualdad:

$$P(x|m) = \int P(x|\,\theta, \mathrm{m})\mathrm{P}(\theta|\mathrm{m})\mathrm{d}\theta$$

Dado el modelo *m* también es posible inferir la para posterior sobre los parámetros del modelo, es decir, $\mathrm{P}(\theta|\mathrm{m})$ y calcular la distribución predictiva $P(x|D, \mathrm{m})$.

Estas cantidades se derivan en analogía exacta con las ecuaciones:

$$P(m|D) = \frac{P(m)P(D|m)}{P(D)} \propto P(m)\prod_{n=1}^{N} P(x_n|m)$$

$$P(x|D) = \sum_{m=1}^{M} P(x|m)P(m|D)$$

Excepto que, en lugar de sumar los posibles modelos, los integramos sobre los parámetros de un modelo en particular. Todas las cantidades clave en el Aprendizaje Automático Bayesiano se deriven directamente de las reglas básicas de la Teoría de Probabilidad. (Ghahramani, 2004)

Vale la pena mencionar algunas formas aproximadas de Aprendizaje Bayesiano. Centrémonos en un modelo particular *m* con parámetros θ y un conjunto de datos observados *D*. Los promedios de la distribución predictiva sobre todos los parámetros posibles ponderados por la parte posterior se representan de la forma:

$$P(x|D, m) = \int P(x|\theta)\mathrm{P}(\theta|\mathrm{D}, \mathrm{m})\mathrm{d}\theta$$

En ciertos casos, puede ser engorroso representar la distribución posterior completa sobre los parámetros, por lo que, en su lugar, se elegirá una estimación puntual sobre los parámetros θ. Una elección natural es elegir el valor del parámetro más probable dados los datos , lo que se conoce como "máximo a posteriori" o MAP. También suele usarse la "máxima verosimilitud" o MLE por su nombre en inglés "maximum likelihood estimation". (Ghahramani, 2004)

La relación con la Estadística va mucho más allá que lo mencionado aquí, sin embargo, no se ahondará más en el tema.

2.3 Casos de Aprendizaje Supervisado

2.3.1 Regresión Lineal - Definición y aclaraciones previas

La Regresión Lineal es uno de los ejemplos más comunes de aplicación de Aprendizaje Supervisado y de acuerdo con (Hastie, Tibshirani, & Friedman, 2008), los modelos lineales se

desarrollaron en gran media en la era de la Estadística antes de la formalización de la computación, pero incluso ahora existen buenas razones para estudiar estos modelos, además de usarlos. Son simples y a menudo proporcionan una descripción adecuada e interpretable de cómo las entradas afectan la salida. A efectos de predicción, a veces pueden superar a los modelos no lineales más sofisticados, especialmente en situaciones con un pequeño número de casos de entrenamiento, bajo la relación señal-ruido o con datos dispersos. Además, los métodos lineales se puede aplicar a las transformaciones de las entradas y esto amplía considerablemente su alcance. Recordemos que, cuando se quiere predecir salidas cuantitativas, hablamos de regresión.

A partir de este momento se usará una notación especial. Se indicará una variable de entrada con el símbolo X . Si X es un vector, se puede acceder a sus componentes mediante los subíndices X_j. Las salidas cuantitativas se indicarán con *Y*, las cualitativas con *G.* Se usarán letras mayúsculas, como X, Y o G, cuando hablemos de los aspectos genéricos de una variable. Los valores observados se escriben en minúsculas, por lo tanto, el último valor observado de X se escribe como x_i (donde x_i es nuevamente un vector). Por el momento, podemos afirmar que la tarea de aprendizaje se representa de la siguiente manera: dado el valor de un vector de entrada *X*, realizar una buena predicción de la salida *Y*, denotada como $\hat{Y}$. Si *Y* toma valores en los reales (IR), entonces también lo hará $\hat{Y}$.

Antes de hablar completamente de la Regresión Lineal, es necesario repasar algunos métodos simples pero poderosos de predicción.

2.3.2 Modelos Lineales y Mínimos Cuadrados

El Modelo Lineal ha sido un pila para la Estadística desde hace más de 30 años y continúa como una de las herramientas más importantes en esta ciencia. Dado un vector de entradas de la forma $X^T = (X_1, X_2, \dots, X_p)$ y podemos predecir la salida *Y* por medio del modelo:

$$\hat{Y} = \hat{\beta_0} + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

El término $\hat{\beta}$ es la intersección, también conocida como sesgo en el aprendizaje automático. A menudo es conveniente incluir la variable constante 1 en *X*, incluir $\hat{\beta}$ en el vector de coeficientes $\hat{\beta}$ y luego escribir el modelo lineal en forma vectorial como un producto interno:

$$\hat{Y} = X^T \hat{\beta}$$

Donde X^T denota un vector o matriz transpuesta. Aquí se está modelando una salida única, por lo que $\hat{Y}$ es de hecho un escalar, aunque bien podría ser un vector, por lo que en ese caso β también sería un vector de coeficientes con relación a las salidas Y. (Hastie, Tibshirani, & Friedman, 2008)

Hay muchos métodos distintos de lograr encajar un modelo lineal en un conjunto de datos de entrenamiento, pero por mucho, el más popular es el método de los Mínimos Cuadrados. En este enfoque, seleccionamos los coeficientes β para minimizar la suma residual de cuadrados:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

$RSS(\beta)$ es una función cuadrática de parámetros y, por lo tanto, su mínimo siempre existe, pero puede que no sea único. La solución es la más fácil de caracterizar en notación matricial y se puede escribir:

$$RSS(\beta) = (y - X\beta)^T (y - X\beta)$$

Donde X es una matriz de $N \times p$ con cada fila como un vector de entrada y y es un vector N de salidas del conjunto de datos de entrenamiento. Al realizar una diferenciación, se obtiene la siguiente ecuación:

$$X^T = (y - X\beta) = 0$$

Si $X^T X$ es no-singular, entonces se obtiene una única solución:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Y el valor ajustado en la entrada x_i sería $\hat{y}_i = \hat{y}(x_i) = x_i^T \hat{\beta}$. Y para un valor arbitrario x_0, la predicción sería $\hat{y}(x_0) = x_0^T \hat{\beta}$. Toda la superficie ajustada se caracteriza por los ρ parámetros $\hat{\beta}$, por lo que parecería que no se necesita un conjunto de datos muy grande para ajustar este modelo. (Hastie, Tibshirani, & Friedman, 2008)

2.3.2.1 Ejemplo

A continuación, se muestra un ejemplo del modelo lineal en un contexto de clasificación.

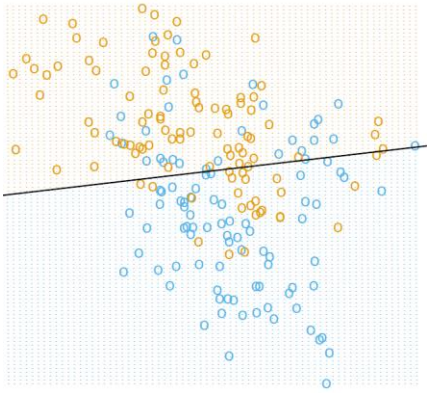


Figura 5. Un ejemplo de clasificación en dos dimensiones, con una codificación de clases en una variable binaria (AZUL= 0, NARANJA=1) y luego, ajustadas por una regresión lineal. En The Elements of Statistical Learning (p. 13), por Hastie, Tibshirani y Friedman.

El conjunto de puntos en IR^2 clasificado como NARANJA corresponde a: $\{x: x^T\beta > 0.5\}$, como se muestra en la figura 5, y las dos clases predichas están separadas por el límite de decisión: $\{x: x^T\hat{\beta} = 0.5\}$, el cual es lineal en este caso. Vemos que para estos datos hay varias clasificaciones erróneas en ambos lados del límite de decisión. Tal vez el modelo lineal sea demasiado rígido, o quizás sean errores inevitables. Es preciso recordar que, estos errores son provenientes de los datos de entrenamiento en sí, y no se ha especificado de dónde provienen los datos construidos. (Hastie, Tibshirani, & Friedman, 2008)

Se consideran los dos siguientes escenarios:

- **Escenario 1:** los datos de entrenamiento en cada clase se generaron a partir de distribuciones gaussianas bivariadas con componentes no correlacionados y diferentes medios.
- **Escenario 2:** los datos de entrenamiento en cada clase provienen de una mezcla de 10 distribuciones gaussianas de baja varianza, con medios individuales distribuidos como gaussianos.

En el primer caso, un límite de decisión lineal es lo mejor que se puede hacer y la optimización es casi óptima; la región de superposición es inevitable, y los datos futuros que se predecirán también se verán afectados por esta superposición. En cuanto al segundo caso, el límite de decisión óptimo es no lineal e inconexo, y como tal es mucho más difícil de obtener.

2.3.3. Método de los vecinos más cercanos (k-vecinos más próximos)

Este método se basa en que, a partir de realizar observaciones en el conjunto de entrenamiento τ más próximo al espacio de entrada a x, para formar $\hat{Y}$. Específicamente, el ajuste del vecino k más cercano a $\hat{Y}$ se define de la siguiente manera:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Donde $N_k(x)$ es el vecindario de x definido por los *k* puntos más cercanos x_i en el conjunto de datos de entrenamiento. La cercanía implica una métrica, que por el momento suponemos que es la distancia euclidiana (es decir, una distancia normal entre dos puntos). Entonces, en otras palabras, encontramos las *k* observaciones con x_i más cercanas a x en el espacio de entrada y promediamos sus respuestas. (Hastie, Tibshirani, & Friedman, 2008)

2.3.3.1 Ejemplos

En la siguiente figura se usan los mismos datos de entrenamiento que en la figura 5, y además se usa como método de ajuste el promedio de 15-vecinos más próximos de la respuesta codificada en binario. Por lo tanto, $\hat{Y}$ es la proporción de NARANJAS en el vecindario y así se asigna la clase NARANJA a $\hat{G}$ si $\hat{Y} > 0.5$ que equivale a un voto mayoritario en el vecindario. Las regiones coloreadas indican todos esos puntos en el espacio de entrada (AZUL o NARANJA) por dicha regla, en este caso se encuentran al evaluar el procedimiento en una cuadrícula fina en el espacio de entrada. Lo primero que llama la atención es que los límites de decisión que separan a las dos opciones son mucho más irregulares y responden a grupos locales donde domina una clase.

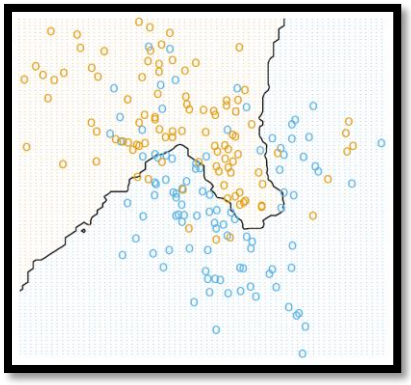


Figura 6. El mismo ejemplo de clasificación que en la Figura 5. El ajuste está realizado por el método de los vecinos próximos (15-vecinos próximos). En The Elements of Statistical Learning (p. 15), por Hastie, Tibshirani y Friedman (2009).

La siguiente figura muestra los resultados para la clasificación de 1-vecino más próximo. A $\hat{Y}$ se le asigna el valor y_l del punto más cercano x_l a x en los datos de entrenamiento. En este caso, las regiones de clasificación pueden calcularse con relativa facilidad y corresponde a una Teselación Voronoi de los datos de prueba. Cada punto x_i tiene un mosaico asociado que delimita la región para la cual es el punto de entrada más cercano. El resultado es que el límite de decisión es aún más irregular que antes. El método de promediar el vecino más cercano k se define exactamente de la misma manera para la regresión de una salida cuantitativa Y, aunque que k tome el valor de 1 sería una opción poco probable. (Hastie, Tibshirani, & Friedman, 2008)

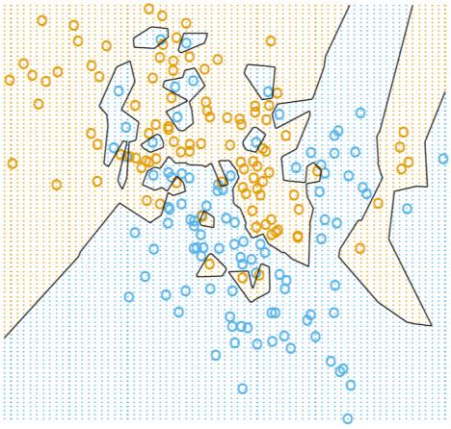


Figura 7. El mismo ejemplo de clasificación que en la Figura 5. El ajuste está realizado por el método de los vecinos próximos (1-vecino próximo). En The Elements of Statistical Learning (p. 16), por Hastie, Tibshirani y Friedman (2009).

En la figura 6 se puede observar que hay muchas menos errores en cuanto a la clasificación (comparando con lo obtenido en la figura 5). Sin embargo, esto no debería ser de mucho consuelo, teniendo en cuenta que en la figura 7, ninguno de los datos de entrenamiento está clasificado erróneamente. Aunque un ejemplo no es de ayuda para llegar a una conclusión, con los ejemplos presentados se podría decir que para los ajustes de k-vecinos más próximos, el error en los datos de entrenamiento debería ser similar a una función creciente de k y siempre será 0 para k=1. (Hastie, Tibshirani, & Friedman, 2008)

2.3.4 Modelos de Regresión Lineal

Como se estableció previamente, se tiene un vector de entrada $X^T = (X_1, X_2, \ldots, X_p)$ y queremos predecir una salida con valor real *Y*. El modelo de Regresión Lineal tiene la siguiente forma:

$$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

El modelo lineal asume que la función de regresión $E(Y|X)$ es lineal o que el modelo lineal es una aproximación razonable. Aquí los β_j son parámetros o coeficientes desconocidos y las variables X_j pueden provenir de distintas fuentes como:

- Entradas cuantitativas
- Transformaciones de entradas cuantitativas como: logaritmos, raíces cuadradas o elevadas al cuadrado.
- Expansiones base como $X_2 = X_1^2$, $X_3 = X_1^3$, etcétera, lo que lleva a una representación polinomial.
- Codificación numérica o "ficticia" de los niveles de entradas cualitativas. Por ejemplo, si G es una entrada de facto de cinco niveles, podríamos crear X_j con $j = 1, \dots, 5$ tal que $X_j = I(G = j)$. Juntos este grupo de X_j representa el efecto de G por un conjunto de constantes dependes del nivel.
- Interacciones entre variables como: $X_3 = X_1 X_2$

No importa la fuente del X_j, el modelo es lineal en sus parámetros. Normalmente se tiene un conjunto de datos de entrenamiento $(x_1, y_1) \dots (x_N, y_N)$ a partir del cual estimar los parámetros β. Cada $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, donde T es un vector de mediciones de características para el caso i. El método de estimación más popular es el de Mínimos Cuadrados, en el que seleccionamos los coeficientes $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ para minimizar la suma residual de cuadrados.

$$RSS\,(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Desde un punto de vista estadístico, este criterio es razonable si las observaciones de entrenamiento (x_i, y_i) representan extracciones aleatorias e independientes de su población. Incluso si los x_i no se escogieron al azar, el criterio sigue siendo válido si los y_i son condicionalmente dependientes dadas las entradas x_i. (Hastie, Tibshirani, & Friedman, 2008)

2.3.5 Clasificación

Dado que un predictor G(x) toma valores de un conjunto discreto G, siempre se puede dividir el espacio de entrada en una colección de regiones etiquetadas según la clasificación; los límites de éstas regiones pueden ser ásperos o suaves, dependiendo de la función de predicción. Para una clase importante de procedimientos, estos límites de decisión son lineales y esto es precisamente lo que se entiende como métodos lineales para la clasificación. (Hastie, Tibshirani, & Friedman, 2008).

Supongamos que existen K clases, por conveniencia etiquetadas de la forma 1,2,3…, K, y el modelo lineal ajustado para la variable de respuesta es: $\widehat{f_k}(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$. El límite de decisión entre la clase k y ℓ es el conjunto de puntos para los que $\widehat{f_k}(x) = \hat{f}_\ell(x)$, es decir, el conjunto $\{ x : (\hat{\beta}_{k0} - \hat{\beta}_{\ell 0}) + (\hat{\beta}_k - \hat{\beta}_\ell)^T x = 0 \}$, un conjunto afín o hiperplano. Como esto es cierto para cualquier par de clases, el espacio de entrada se divide en regiones de clasificación constante, con límites de decisión hiperplanar por partes. Este enfoque de regresión es un miembro de una clase de métodos que modelan las funciones discriminantes $\delta k(x)$ para cada clase y luego asignan x a la clase con el mayor valor para su función discriminante. Los métodos que modelan las probabilidades posteriores $\Pr(G = k|X = x)$ también están en esta clase. Claramente si $\delta k(x)$ o $\Pr(G = k|X = x)$ son lineales en x, entonces los límites de decisión serán lineales también. (Hastie, Tibshirani, & Friedman, 2008)

En realidad, todo lo que se necesita es que alguna transformación monótona de $\delta k(x)$ o $\Pr(G = k|X = x)$ sea lineal para que los límites de decisión también lo sean. Por ejemplo, si hay dos clases, un modelo popular para las probabilidades posteriores es:

$$\Pr(\mathrm{G} = 1|\mathrm{X} = \mathrm{x}) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$\Pr(\mathrm{G} = 2|\mathrm{X} = \mathrm{x}) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$

En este caso, la transformación monótona es la transformación *logit:* $\log[\frac{p}{(1-p)}]$ por lo que se ve entonces:

$$log \frac{\Pr(\mathrm{G} = 1|\mathrm{X} = \mathrm{x})}{\Pr(\mathrm{G} = 2|\mathrm{X} = \mathrm{x})} = \beta_0 + \beta^T x$$

El límite de decisión es el conjunto de puntos para los cuales las probabilidades de registro son cero y este es un hiperplano definido por $\{x|\ \beta_0 + \beta^T x = 0)$. Se discuten dos métodos muy populares pero diferentes que resultan en probabilidades logísticas o *logit*: análisis discriminante lineal y regresión logística lineal. Aunque difieren en su derivación, la diferencia esencial entre ellos está en la forma en que la función lineal se ajusta a los datos de entrenamiento. Un enfoque más directo es modela explícitamente los límites entre las clases como lineales. Para un problema de dos clases en un espacio de entrada p-dimensional, esto equivale a modelar el límite de decisión como un hiperplano; en otras palabras, un vector normal y un punto de corte. Veremos dos métodos que buscan explícitamente "separar los hiperplanos". El primero es el conocido modelo de percepción de Rosenblatt, con un algoritmo que encuentra un hiperplano de separación en los datos de entrenamiento si este existe. El segundo método realizado por Vapnik en 1996, encuentra un hiperplano que se separa de manera óptima si existe, de lo contrario, encuentra un hiperplano que minimiza cierta medida de superposición en los datos de entrenamiento. (Hastie, Tibshirani, & Friedman, 2008)

2.5 Análisis Clúster

2.5.1 Definición

Así como vimos ejemplos de Aprendizaje Supervisado con los temas de Regresión Lineal y Clasificación, en este caso se profundizará en el tema de Clustering.

El Análisis Clustering o también llamado "Segmentación de Datos" debe su nombre a la palabra "clúster" de origen inglés y que significa "conglomerado". Los objetivos de este análisis son variados, pero todos se relacionan con la agrupación o segmentación de una colección de objetos en subconjuntos o "clústeres" , de modo que aquellos dentro de cada agrupación están más estrechamente relacionados entre sí que los objetos asignados a diferentes agrupaciones. Un objeto puede ser describo por un conjunto de medidas o por su relación con otros objetos. Además, el objetivo a veces es organizar los clústeres en una jerarquía natural. Esto implica agrupar sucesivamente los clústeres en sí mismos para que, en cada nivel de jerarquía, los grupos dentro del mismo grupo sean más similares entre sí que a los de grupos diferentes. El Análisis Clúster también se usa para formar estadísticas descriptivas para determinar si los datos consisten o no en un conjunto de subgrupos distintos, cada grupo representa objetos con propiedades sustancialmente diferentes. Este último objetivo requiere una evaluación del grado

de diferencia entre los objetos asignados a los respectivos grupos. La noción del grado de similitud o disimilitud entre los objetos individuales que se agrupan es fundamental para todos los objetivos del Análisis Clúster. Un método de agrupación interna busca la integración de los objetos según su definición, lo cual solamente puede venir de las consideraciones de la materia. Esta situación puede verse como algo similar a la especificación de una función de pérdida en los problemas de predicción del Aprendizaje Supervisado. (Hastie, Tibshirani, & Friedman, 2008)

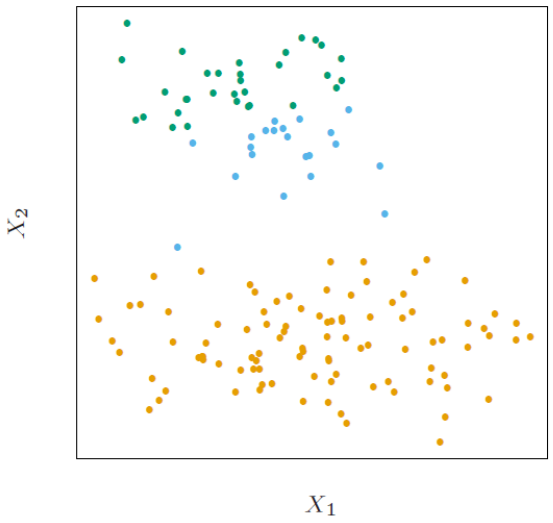


Figura 8. Datos simulados en el plano, agrupados en tres clases (representadas en color naranja, azul y verde) por el algoritmo de Clustering "K-medias". En The Elements of Statistical Learning (p. 502), por Hastie, Tibshirani y Friedman (2009).

La Figura 8 muestra algunos datos simulados que están agrupados en tres categorías por medio del popular algoritmo "K-medias". En este caso, dos de los clústeres no están bien separados, por lo que la "segmentación" describe con mayor precisión la parte de este proceso que el "agrupamiento" en sí mismo. El agrupamiento "K-medias" comienza con conjeturas para los tres centros del clúster. Luego se alternan los siguientes pasos hasta la convergencia:

- Para cada punto de datos, se identifica el centro del clúster más cercano (en distancia euclidiana).
- Cada centro del clúster se reemplaza por el promedio de coordenadas de todos los puntos de datos que están más cerca de él. (Hastie, Tibshirani, & Friedman, 2008)

2.5.2 Matrices de Proximidad

A veces, los datos se representan directamente en términos de proximidad entre pares de objetos. Estas puede ser similitudes o diferencias. Por ejemplo, en experimentos sociales se suele pedir a los participantes que juzguen en qué medida ciertos objetos se diferencian entre sí. Las diferencias se pueden calcular promediando la recopilación de dichos juicios. Este tipo de datos se puede representar mediante una matriz D de dimensiones $N \times N$, donde N es el número de objetos y cada elemento d_{ii} guarda la proximidad entre dos objetos. Esta matriz es dada como entrada en los algoritmos de clústering. La mayoría de algoritmos suponen una matriz de diferencias con entradas no negativas y cero elementos diagonales: $d_{ii} = 0; I = 0,1,2, \dots, N$. Si los datos originales se recopilaron como similitudes, se puede utilizar una función monótona adecuada de disminución para convertirlos en "diferentes". Además, la mayoría de los algoritmos asumen matrices de similitud simétrica, por lo que si la matriz D original no es simétrica debe reemplazarse por $\frac{D \times D^T}{2}$. (Hastie, Tibshirani, & Friedman, 2008)

2.5.3 Algoritmos Clústering

El objetivo del Análisis Clúster es dividir las observaciones en grupos o clústeres de modo que las diferencias entre los pares asignados al mismo clúster tiendan a ser más pequeñas que las de los clústeres diferentes. Los algoritmos clústering se dividen en tres tipos distintos

- **Algoritmos combinatorios:** funcionan directamente en los datos observados sin una referencia directa a un modelo de probabilidad subyacente.
- **Modelado de Mezclas:** supone que los datos son una muestra independiente e idénticamente distribuida de una población descrita por una función de densidad de probabilidad. Esta función de densidad se caracteriza por un modelo parametrizado que se considera una "mezcla" de funciones de densidad de componentes, donde cada densidad de componentes describe uno de los grupos. Este modelo luego se ajusta a los datos mediante Máxima Verosimilitud o por los enfoques bayesianos correspondientes.
- **Buscadores de Modo o "Cazadores de Golpes":** adoptan una perspectiva no paramétrica, intentando estimar directamente los distintos modos de la función de densidad de probabilidad. Las observaciones más cercanas a cada modo respetivo definen los clústeres individuales. (Hastie, Tibshirani, & Friedman, 2008)

Uno de los algoritmos mencionados previamente, el algoritmo "K-medias", es uno de los más conocidos en el mundo de las técnicas de clústering; el mismo se describe a continuación:

Tal y como mencionan Tan, Steinbach, & Kumar (2006), primero se escogen K centroides (siendo un centroide la media de un grupo de puntos), donde K es un parámetro especificado por el usuario y representa el número de clústeres deseados. Cada punto es entonces asignado al centroide más cercano. Cada colección de puntos asignados a un centroide es un clúster. El centroide de cada clúster es actualizado en base a los puntos asignados al mismo. Se repite la asignación y la actualización de los clústeres hasta que ningún punto altere al clúster, o lo que es lo mismo, hasta que los centroides se mantengan iguales.

2.6 Análisis del Marco Teórico Incluido

2.6.1 El Aprendizaje de Máquina en Minería de Datos.

Durante este capítulo se han revisado algunos puntos acerca del Aprendizaje de Máquina que es utilizado por el software dedicado a la Minería de Datos. Los algoritmos de Aprendizaje de Máquina se emplean con diferentes objetivos a la hora de preparar los datos, modelarlos o analizarlos y de esta forma encontrar patrones o generar predicciones que aporten a la estructuración de conocimiento o también realizar predicciones.

Como puede resultar obvio, la elección de una técnica en específico deberá pasar por algún tipo de análisis en el cual se sobrepongan las ventajas para cada situación, es decir, no se puede decir que existe algún algoritmo que es siempre mejor para su aplicación con Minería de Datos. Las aplicaciones de Minería de Datos en campos industriales o comerciales resultan ser especialmente desafiantes en términos de los requisitos impuestos a los procedimientos de aprendizaje, por ejemplo, los conjuntos de datos a menudo resultan ser muy grandes en términos de número de observaciones y número de variables medidas em cada una de ellas. Por lo tanto, las consideraciones computacionales juegan un papel importante. Además, como se habla de ejemplos prácticos de la vida real, se tiende a tratar con datos desorganizados donde las entradas suelen ser una mezcla de variables cuantitativas, binarias y categóricas (con muchos niveles); sin mencionar que es usual tener valores faltantes, no existen muchas observaciones completas, etcétera.

En las aplicaciones de Minería de Datos, generalmente solo una pequeña fracción de la gran cantidad de variables predictoras que se han incluido en el análisis son realmente relevantes

para la predicción. Además, a diferencia de muchas aplicaciones como el simple reconocimiento de patrones, es poco usual que exista un conocimiento de dominio confiable para ayudar a crear características especialmente relevantes y/o eliminar las irrelevantes (que son especialmente perjudiciales al degradar de manera drástica el rendimiento en muchos métodos). Estas aplicaciones suelen requerir modelos interpretables; no es suficiente con producir predicciones y es deseable contar con información que proporcione comprensión cualitativa entre de la relación de entrada y el valor de respuesta previsto. Por lo tanto, algunos de los métodos más populares y efectivos en entornos puramente predictivos como las Redes Neuronales, pueden resultar mucho menos útiles en Minería de Datos.

Estos requisitos de velocidad, interpretabilidad y la naturaleza desordenada de los datos limitan drásticamente la utilidad de la mayoría de los procedimientos de aprendizaje como métodos estándar para la Minería de Datos. Un método "listo para usar" es uno que se puede aplicar directamente a los datos sin requerir mucho procesamiento previo de los mismos, o un ajuste cuidadosos del procedimiento de aprendizaje. Según la opinión de Hastie, Tibshirani y Friedman, (2009), a pesar de que los métodos más populares son los citados en este trabajo (Clústering, Clasificación, Regresión), existe un método adicional que está mucho más cerca de cumplir con los requisitos para actuar como un procedimiento estándar para la Minería de Datos: los árboles de decisión, esto aunque los mismos autores mencionan que este método posee una gran desventaja: la inexactitud y es que los árboles de decisión rara vez proveen una predicción aceptable en comparación a lo que se podría lograr con los datos disponibles.

2.6.2 La importancia de los distintos métodos tratados.

Una vez analizados los distintos métodos tanto de Aprendizaje Supervisado como No Supervisado se puede notar la importancia de conocer su funcionamiento a través de las Matemáticas de modo que se conozca totalmente hacia qué problema se apegan. Por ejemplo, cuando se tratan problemas de Clasificación se espera tener mucha influencia de las variables cualitativas del estudio para poder generar justamente una lógica detrás de las clasificaciones otorgadas, de igual modo sucede con el tema de Clustering, que es un método de Aprendizaje No Supervisado. El estudio matemático y estadístico que se desprende de estos temas se ha tratado de resumir de modo que el lector tenga un conocimiento general de lo que espera conseguir al aplicar las herramientas de software en un futuro análisis, mas no busca centrarse en los detalles de cada uno de los métodos tratados. La elección de los métodos se enfocará en las facilidades

en temas de programación y el propio entorno de las herramientas, aunque los métodos más populares como Regresión Lineal usualmente están presentes en las aplicaciones de Minería de Datos.

La idea general de la revisión de las técnicas, métodos y algoritmos planteados es la de tener una visión más precisa de cómo funcionan las herramientas de software, su estructura y lo que se puede realizar en ellas. No se deben confundir los objetivos del presente trabajo con el desarrollo de una herramienta de software nueva de Minería de Datos, por lo que, se deben encontrar productos de software y comprender lo que se puede realizar con los mismos.

3. Obtención de Datos y Elección de Herramientas.

3.1 Definición de los repositorios de datos a usar desde el Ministerio de Educación.

Para la respectiva aplicación de Data Mining, se usarán datos públicos ofrecidos por el Ministerio de Educación de la República del Ecuador, estos datos se pueden encontrar en la página web de esta institución, en la sección de Estadística Educativas. Los datos se encuentran clasificados de acuerdo al año lectivo y separados por género, edad, región, ciudad, entre otros factores, de los estudiantes de Educación Básica y Bachillerato. Estos datos se catalogan como "AMIE" (Archivo Maestro de Instituciones Educativas) y son estadísticas recogidas desde el año 2009 hasta el inicio del año lectivo 2017 (lamentablemente el fin del año lectivo 2017 y el inicio del 2018-2019 no están disponibles).

Considerando que, la cantidad de datos es amplia y que el conglomerado de datos abarca casi una década, se opta por no tomar en consideración a los datos catalogados bajo el nombre de "SINEC", que son las estadísticas que abarcan desde el año lectivo 1993-1994 hasta el 2006-2007.

El objetivo principal de este trabajo, se recalca, será el de generar predicciones que puedan ser útiles para estudios futuros; a la par que se encontrarán patrones como los que se mencionan a continuación:

- Que la tendencia en la región Sierra del Ecuador sea que el ingreso de niños a la Educación Básica, en instituciones fiscales, se dé entre los 3 y 4 años.
- Que exista una tendencia favorable a tener más mujeres en instituciones particulares en la provincia de Pichincha.
- Que exista una relación independiente entre el número de docentes en instituciones fiscales y el número de alumnos de cada una de dichas instituciones.
- Etcétera.

Debido a la amplia cantidad de los datos, un análisis informal no se precisa como suficiente y es por lo que se requiere a la Minería de Datos como herramienta principal para obtener patrones como los previamente descritos.

Los datos están clasificados en archivos .xlsx distintos, dos por cada período lectivo indicando el inicio y el fin de dichos períodos. La formalización de los datos no necesariamente

implicará la creación de bases de datos al uso, puesto que existen herramientas que aceptan como entrada un archivo .xlsx o un archivo .csv cuya conversión será realizada cuando sea requerida.

Por otro lado, lo que sí se realizará con antelación será el agrupamiento de datos de acuerdo a factores más generales como, por ejemplo, catalogar los datos de todas las provincias de la Sierra en un mismo grupo (igualmente con las demás regiones del país) y también se agruparán los datos en un mismo archivo para observar los diferentes resultados. La factibilidad del uso de estos datos entrará en juego cuando se empleen las herramientas de Minería, en el caso de que los datos, en su estado puro (es decir, tal y como se los encontró en la fuente) no sean factibles de usar, serán sometidos a los distintos procesos que sean requeridos para que cumplan con su propósito.

Los métodos, herramientas y procesos usados para la obtención de los datos que se emplean en el presente trabajo son producto, en su totalidad, del Ministerio de Educación por lo que, su veracidad, precisión, limitaciones, omisiones o posibles contradicciones son responsabilidad de esta institución y se remite al AMIE como fuente principal para solventar cualquier duda con respecto a este tema.

La estructura de los archivos descargados correspondientes se puede apreciar en la figura a continuación:

Ministerio de Educación

Elaboración: Dirección de Análisis e Información Educativa (DNAIE) / Coordinación General de Planificación (CGP) / Ministerio de Educación (MinEduc)
Fuente: Archivo Maestro de Instituciones Educativas (AMIE) Período 2017-2018 inicio

	UBICACIÓN POLITICA ADMNISTRATIVA				
Periodo	Provincia	Cod_Provincia	Cantón	Cod_Cantón	Parroquia
2017-2018 Inicio	AZUAY	01	CUENCA	0101	BAÑOS
2017-2018 Inicio	AZUAY	01	CUENCA	0101	CHAUCHA
2017-2018 Inicio	AZUAY	01	CUENCA	0101	QUINGEO
2017-2018 Inicio	AZUAY	01	CUENCA	0101	QUINGEO
2017-2018 Inicio	AZUAY	01	CUENCA	0101	RICAURTE
2017-2018 Inicio	AZUAY	01	CUENCA	0101	SANT ANA

Figura 9. Conjunto de Datos de Instituciones Educativas en el Ecuador, por Ministerio de Educación.

Cada archivo está catalogado en el período correspondiente y además de los datos, posee un diccionario con la explicación correspondiente de las columnas.

3.2 Análisis previo de los datos elegidos.

3.2.1 Descripción de la estructura de los archivos

Los datos que se pueden obtener de los archivos del AMIE se encuentran clasificados bajo una serie de categorías (como se indica en la Figura 9), de las que se resaltan las siguientes:

- Período: corresponde al período lectivo.
- Provincia: el nombre de la provincia respectiva.
- Cantón: el nombre del cantón respectivo.
- Parroquia: el nombre de la parroquia respectiva.
- Nombre de la Institución: el nombre completo de la institución educativa.
- Dirección de la Institución: la dirección general de la institución educativa.
- Si se encuentra escolarizada o no
- Tipo de Educación: que bien puede ser Regular, Popular Permanente, Artística, entre otras.
- Nivel de Educación: que bien puede ser Inicial, Educación Básica, Bachillerato.
- Sostenimiento: de dónde provienen los fondos de la institución y que bien puede ser Particular, Fiscal, Municipal, entre otros
- Régimen Escolar: como es de conocimiento general, existen dos regímenes para la educación básica y secundaria: Sierra y Costa.
- Modalidad (presencial, semipresencial, a distancia)
- Jornada
- Docentes: el número de docentes subdividido en las siguientes categorías:
- Docentes femeninos
- Docentes masculinos
- Administrativos: el número de empleados de la institución, no relacionados con la docencia y subdividido en las siguientes categorías:
- Mujeres
- Hombres
- Estudiantes: el número de estudiantes por género.
- Mujeres
- Hombres

- Número de estudiantes por cada nivel de educación: son los datos de manera más detallada de cada uno de los niveles académicos y se subdividen en las siguientes categorías:
- Sexo: el género de los estudiantes, es la categoría general.
- Promovidos: los estudiantes que pudieron acceder al siguiente año académico, se subdividen en hombres y mujeres.
- No promovidos: los estudiantes que no pudieron acceder al siguiente año académico, se subdividen en hombres y mujeres.
- Abandono: los estudiantes que desertaron de la educación (no se especifican causas), , se subdividen en hombres y mujeres.
- No actualizados: una categoría de respaldo por si se tienen dudas; las instituciones pueden llenar este apartado con alguna irregularidad en caso de ser requerida. Igualmente presentan la subdivisión de hombres y mujeres. Esta categoría podría no tomarse en cuenta en futuros análisis debido al bajo número que a simple vista se nota que posee.

3.2.2 Preparación de los Datos

Tal y como menciona Microsoft en la documentación en línea de SQL Server, los datos que se procedan para la aplicación de Minería de Datos no necesitan agruparse en estructuras de datos complejas como un cubo OLAP, ni siquiera en una base de datos relacional, aunque es evidente que existirían ventajas de acuerdo a la herramienta de software que se emplee. Por lo tanto, la Minería de Datos se puede aplicar a cualquier conjunto de datos con algún orden u agrupamiento lógico, como bien podría ser un archivo de Excel.

Uno de los beneficios de usar los archivos procedentes del AMIE es la homogeneidad de los mismos, específicamente en las siguientes condiciones:

- Todos los archivos tienen la extensión .xlsx que corresponden a hojas de cálculo de Excel (en sus versiones más actuales).
- Todos los archivos contienen las mismas columnas (como se pudo apreciar previamente).
- Dadas las condiciones previamente planteadas, se procede a crear nuevos archivos con las características que se desean. Además, los estudiantes que se rijan a

programas especiales como los que se citan a continuación, no serán tomados en cuenta para los posteriores análisis:

- Estudiantes Artesanal
- Estudiantes Básico Acelerado.
- Estudiantes Alfabetización.
- Estudiantes Post-Básico.
- Estudiantes "Desconoce".
- Estudiantes No Escolarizado.
- Estudiantes de preescolar (menores de 3 años, 3 y 4 años).

Los archivos estarán almacenados en una carpeta exclusiva. Los nombres para los archivos originales serán los siguientes:

- "Registros-Adminstrativos-20XX-20XX-fin-O".xlsx

En cuanto a las actualizaciones para los análisis, los archivos tendrán los siguientes nombres:

- "Registros-Adminstrativos-20XX-20XX-fin-A".xlsx

Para cada período lectivo existen dos archivos: uno con los datos de cuando se inicia el período y otro cuando se finaliza dicho período; para los análisis correspondientes solamente se trabajará con los datos de fin de período. Como se explicó previamente, el período lectivo 2017-2018 es especial pues no contiene datos de fín de período, por lo que no se lo tomará en cuenta. Los archivos serán organizados por periodo lectivo para más comodidad y se descartarán todos los datos que no cumplan con las características mencionadas.

Como último paso, se eliminarán las columnas con los "totales", como es el caso en profesores y administrativos, y todas las columnas que agrupan a los estudiantes; esto debido a que resulta redundante tener los datos agrupados y desglosados. Este proceso será aplicado para todos los archivos comprendidos en el AMIE con el fin de servir de primer filtro para la posterior aplicación de Minería de Datos.

3.2.2.1 Ajuste de datos en cero

A pesar de que se aplicó un filtro inicial eliminando ciertas columnas con datos que no aportarían mucho en futuros análisis, los archivos del AMIE podría aún contar con columnas completas con valores en cero, que bien podrían ser descartados. Es evidente que no se podrá realizar esta revisión manualmente debido a la gran cantidad de datos de cada archivo. En el caso de que uno de los archivos contenga una columna con todos sus valores en cero, pero en los demás archivos esta misma columna contenga valores distintos, la columna no será eliminada.

Como se están tratando únicamente datos cuantitativos, se puede comprobar si una columna contiene o no valores distintos a cero mediante una suma de todos sus datos, estos se puede hacer con Excel y este valor nos dirá si es precio eliminar dicha columna.

Revisión de valores de todas las columnas

Período 2009-2010

	ESTUDIANTES PRIMER AÑO BÁSICA																ESTUDIANTES SEGUNDO
	DOCENTES		ADMINISTRATIVOS		ESTUDIANTES		SEXO		PROMOVIDOS		NO PROMOVIDOS	ABANDONO		NO ACTUALIZADOS		SEXO		PROMOVIDOS		NO PROMOVID
	Docentes Femenino	Docentes Masculino	Administrativos Femenino	Administrativos Masculino	Estudiantes Femenino	Estudiantes Masculino	EstudiantesFemeninoPrimerAñoEGB	EstudiantesMasculinoPrimerAñoEGB	EstudiantesFemeninoPromovidosPrimerAñoEGB	EstudiantesMasculinoPromovidosPrimerAñoEGB	EstudiantesFemeninoNoPromovidosPrimerAñoEGB	EstudiantesMadculinoNoPromovidosPrimerAñoEGB	EstudiantesFemeninoDesertoresPrimerAñoEGB	EstudiantesMasculinoDesertoresPrimerAñoEGB	EstudiantesFemeninoNoActualizadoPrimerAñoEGB	EstudiantesMasculinoNoActualizadoPrimerAñoEGB	EstudiantesFemeninoSegundoAñoEGB	EstudiantesMasculinoSegundoAñoEGB	EstudiantesFemeninoPromovidosSegundoAñoEGB	EstudiantesMasculinoPromovidosSegundoAñoEGB	EstudiantesFemeninoNoPromovidosSegundoAñoEGB	Estu Masc oPro sSegu o
Totales	153350	76341	22200	15497	2064052	2089130	146296	150517	132023	135207	3685	3915	5383	5833	5205	5562	187777	197545	168799	174655	6857	

Figura 10. Revisión de resultados para encontrar posibles columnas en cero. (Elaboración Propia)

En la Figura 10 se muestran los valores totales de algunas columnas. Tras realizar este proceso a cada uno de los archivos, se encontró que ninguna de las columnas que se seleccionaron contenían todos sus valores en cero, es decir, no amerita la eliminación de las mismas debido a que podrán resultar útiles estos valores para la búsqueda de patrones y/o predicciones. En cuanto al proceso de preparación de datos, se podría decir que no existe nada más por completar, sin embargo, aún se presentarán algunos añadidos como gráficas para comprender más todos los datos con los que se pretende trabajar.

Por lo que, los datos de planteles educativos elegidos son:

- Planteles con tipo de educación: primaria, secundaria o bachillerato. Aquellos planteles que tengan el "bachillerato" como tipo contendrán los previos tipos también, es decir, si un plantel educativo aparece en el conjunto de datos como Bachillerato, querrá decir que tiene estudiantes desde primero de básica hasta tercero de bachillerato.
- Lo anterior implica que los demás tipos de educación, incluyendo educación inicial quedan descartados para futuros análisis.

- Debido a la reestructuración llevada a cabo, puede que algunos planteles pierdan estudiantes de niveles de educación inicial (niños menores de 4 años) al quedar desafectados del grupo final de datos.
- Se reestructuran los tipos de Jornada para quedar de la siguiente manera: Matutina, Vespertina, Nocturna, MatVes (Matutina y Vespertina), MatVesNoc (Matutina, Vespertina y Nocturna), MatNoc (Matutina y Nocturna), VesNoC (Vespertina y Nocturna).

3.2.3 Gráficas Preliminares

Debido a que están a disposición archivos que comprenden un buen puñado de años, es factible presentar gráficas que muestren la evolución de ciertos datos en el tiempo, con el fin de tener una idea general que pueda aportar previo a la aplicación de Minería de Datos. A continuación, se presentan un conjunto de gráficos con respecto a tres importantes variables como son los acumulados de: profesores, administrativos y estudiantes divididos entre hombres y mujeres.

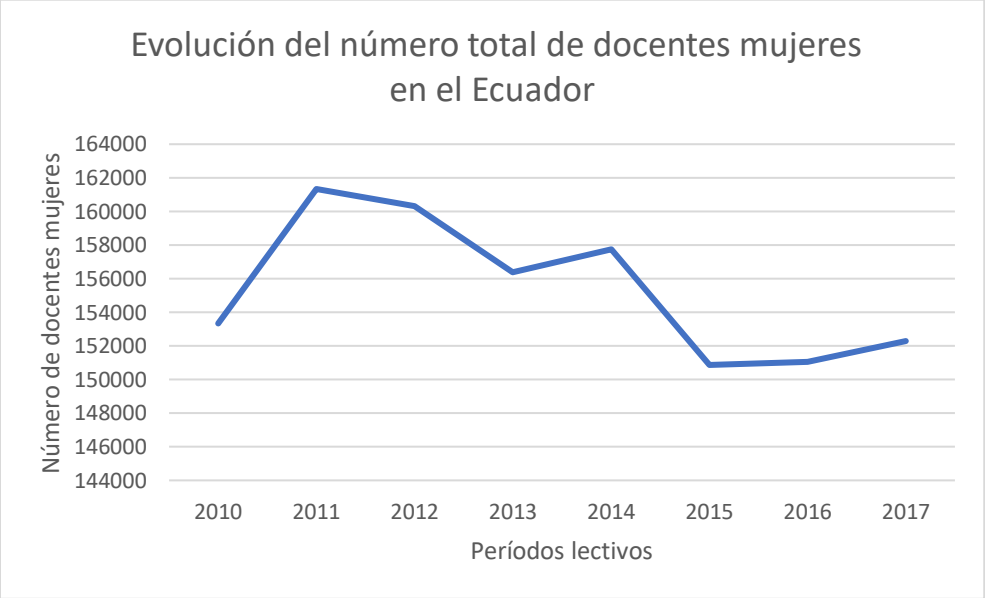


Figura 11. Evolución del número total de docentes mujeres en el Ecuador (Elaboración Propia)

En la Figura 11 podemos evidenciar cómo existe una variación considerable del número de docentes mujeres entre cada finalización de los períodos lectivos, siendo el máximo obtenido en el año 2011 y el mínimo en 2015. Sin embargo, los datos indican que existen más docentes mujeres que hombres, en todos los períodos lectivos analizados, siendo que el máximo número

de docentes mujeres alcanzado en estos años, es casi el doble del máximo número de docentes hombres.

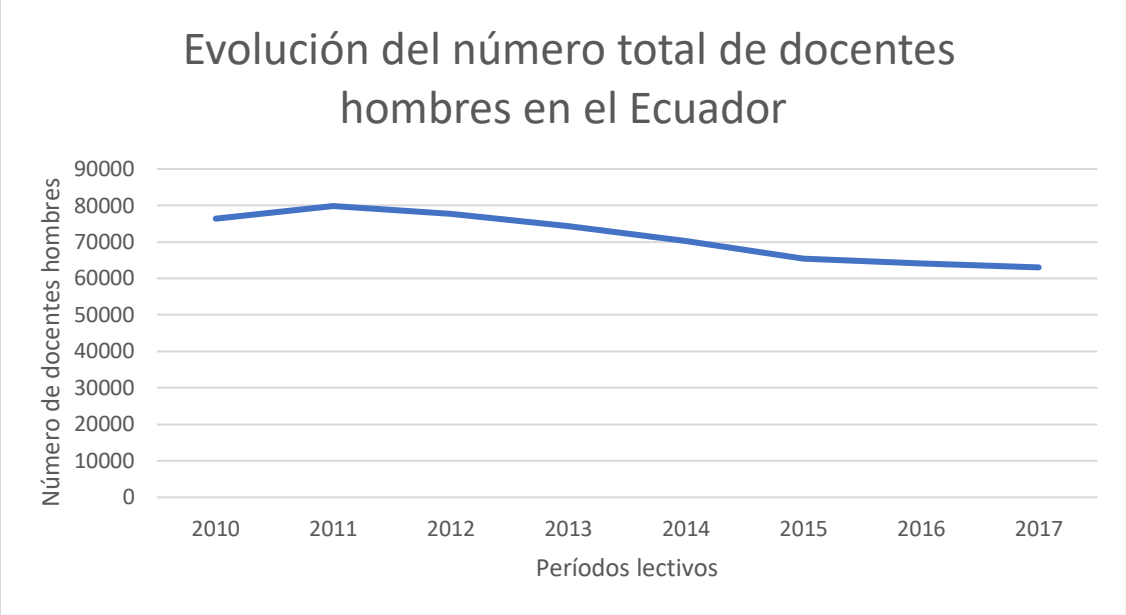


Figura 12. Evolución del número total de docentes hombres en el Ecuador. (Elaboración Propia)

En la Figura 12 se puede corroborar la información mencionada previamente, en donde tenemos una constancia muy marcada en estos años, con un promedio de alrededor de 71368 docentes hombres en el país. Un dato para resaltar es que, al igual que ocurre con los docentes mujeres, el número de hombres disminuye desde el año 2011, para no volver a subir en períodos posteriores.

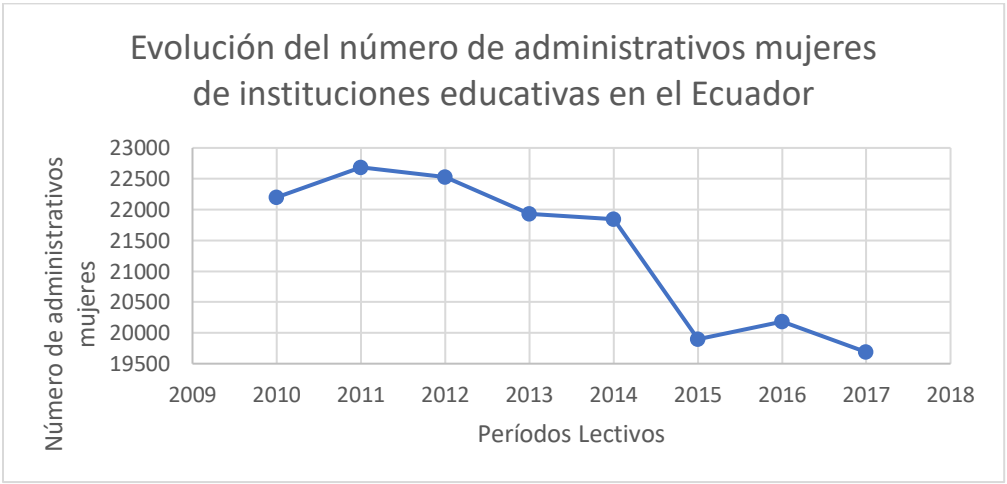


Figura 13. Evolución del número de administrativos mujeres de instituciones educativas en el Ecuador. (Elaboración Propia)

En la Figura 13 se muestra el número del personal administrativo en instituciones educativas en el Ecuador. A pesar de que podría pensarse que es un dato sin importancia para obtener patrones y/o predicciones importantes en el mundo de la educación nacional, es menester de todos recordar que una institución educativa también necesita ser administrada y no sale a flote únicamente con los docentes. El número de personas administrativas de sexo femenino ronda los 21000 individuos como promedio (aproximadamente). El número presenta una variación fuerte en el año 2014 en el cual se da una disminución de casi 2000 personas para el siguiente período lectivo.

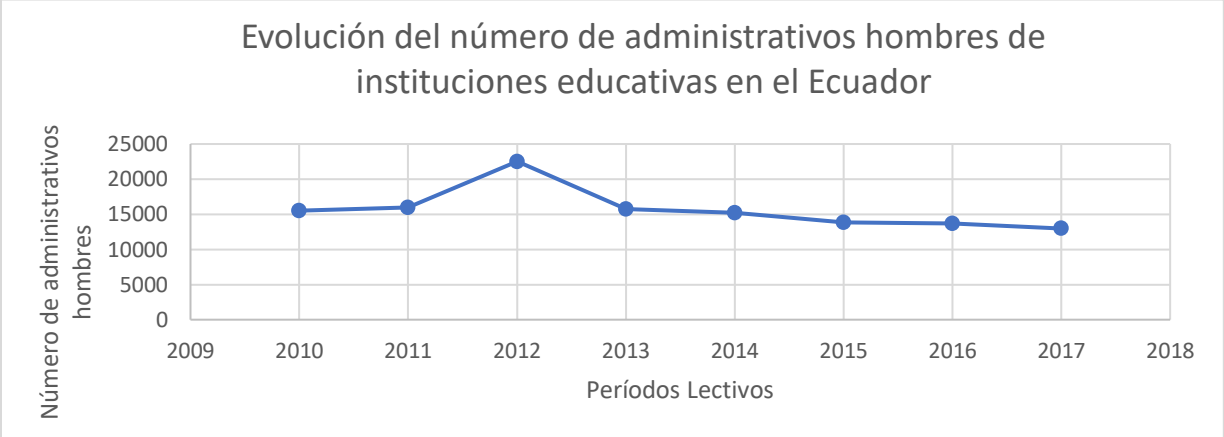


Figura 14. Evolución del número de administrativos hombres de instituciones educativas en el Ecuador (Elaboración Propia)

La Figura 14 muestra los número de administrativos hombres, que parece repetir el caso de docentes hombres con datos similares período tras período, con un promedio de 15000 personas y teniendo un pico en el año 2012.

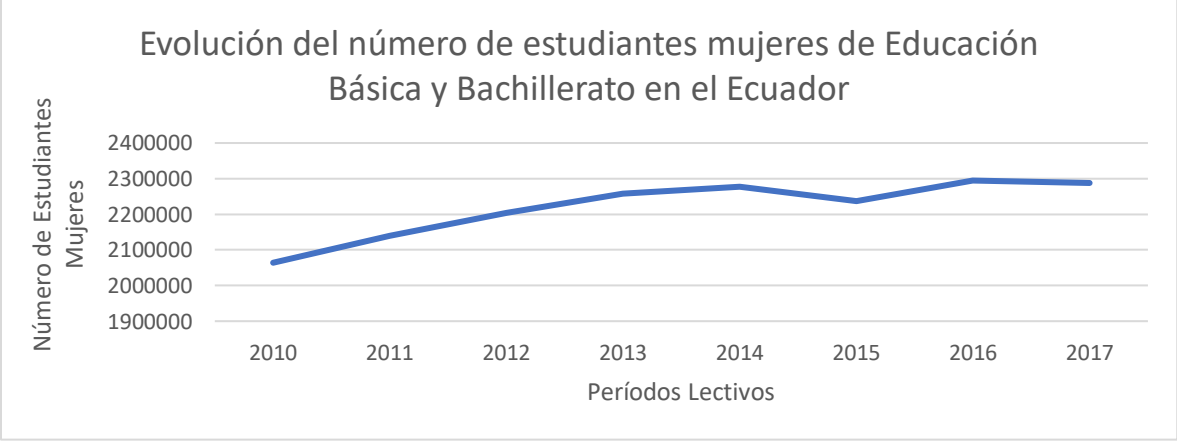


Figura 15. Evolución del número de estudiantes mujeres de Educación Básica y Bachillerato en el Ecuador (Elaboración Propia)

La Figura 15 muestra los datos de estudiantes mujeres en el Ecuador, con un evidente crecimiento desde el año 2010 hasta el 2014, para después mostrar un leve descenso en el 2015 y posteriormente vuelve a elevarse. Estos son los datos más amplios con un promedio de 2 millones doscientos mil estudiantes.

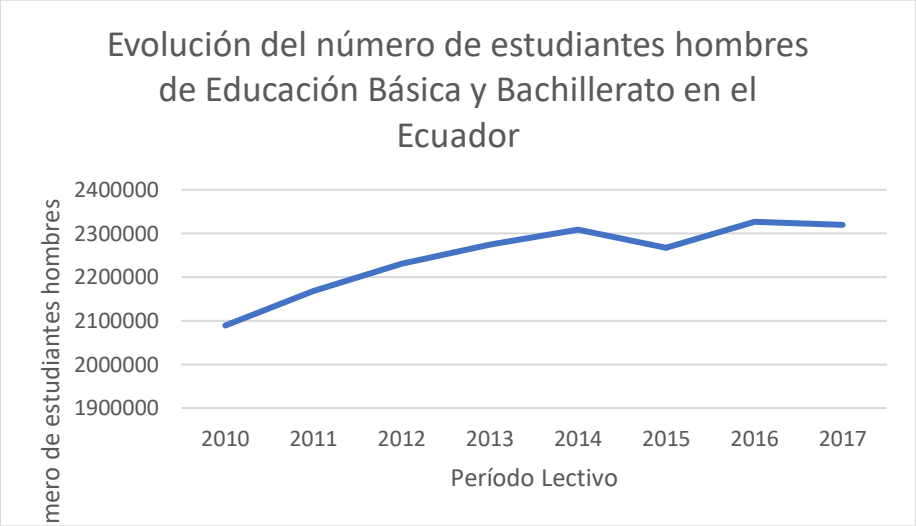


Figura 16. Evolución del número de estudiantes hombres de Educación Básica y Bachillerato en el Ecuador (Elaboración Propia)

El gráfico 6 muestra los datos de estudiantes hombres, en donde se aprecia una tendencia casi exactamente similar que los datos de estudiantes mujeres. El promedio de estudiantes hombres es apenas 28000 personas más que las mujeres (aproximadamente).

Estas primeras gráficas nos muestran ciertas características de los datos en el transcurso del tiempo, como ciertas subidas o bajadas considerables de números en un determinado período o situaciones opuestas entre hombres y mujeres. Una vez se analicen estos datos a profundidad, se espera obtener patrones y/o predicciones que permitan complementar la información levantaba inicialmente, así como resolver dudas adicionales que puedan surgir al seguir analizando este conjunto de datos.

3.3 Herramientas de Minería de Datos elegidas

3.3.1 R y R Studio

R es un sistema que se enfoca en la computación estadística y generación de gráficos. En sí mismo, se puede considerar a R como un lenguaje de programación, aunque en realidad en un sistema que también provee gráficos de alto nivel, interfaces para otros lenguajes y facilidades de

depuración. La creación de este lenguaje se remonta a los años 80 y cuenta con una sintaxis simple en donde la mayoría de usuarios no necesitan aprender realmente la estructura entera del lenguaje para poder desarrollar aplicaciones funcionales. Para trabajar con R, se ha elegido un IDE reconocido por ser gratis y de código abierto, llamado R Studio o RStudio. (R Core Team, 2018).

Este IDE fue desarrollado por el ingeniero de software Joseph J. Allaire, reconocido internacionalmente por haber creado el lenguaje de programación ColdFusion. Una de las ventajas de trabajar con este IDE y con R es la gran cantidad de librerías que son gratuitas de usar y pueden ser fácilmente añadidas a cualquier proyecto desarrollado con este software. Existen dos versiones: la de escritorio (con la cual se trabajará) y una versión para navegadores web que funciona a través de un servidor Linux. Las versiones gratuitas no cuentan con ninguna restricción mayor para trabajar en proyectos no tan extensos. (RStudio, 2018)

Tabla 1. R y RStudio

Foco de Atención	Centrado en la creación de gráficos y la computación estadística
Características principales	Gratis y de código abierto. Facilidad en sintaxis, muchas librerías disponibles gratuitamente. Disponible en muchas plataformas, muy usado para Data Science.
Lenguaje	Escrito en JAVA, C++ y Javascript.
Interacción con el usuario	Por medio de GUI.
Licenciamiento	Affero General Public License v3/ Propietario (versiones profesionales)
Última versión estable	8 de Abril del 2019.

Principales características de la herramienta RStudio (Elaboración Propia)

3.3.2 Orange

Es un conjunto de herramientas open source para visualización de datos, Aprendizaje de Máquina y Minería de Datos. Cuenta con una interfaz de programación visual para el análisis exploratorio de datos y la visualización interactiva de datos. Puede usarse como una librería en Python.

En Orange, los componentes son denominados "widgets" y existen de varias complejidades como la simple visualización de datos hasta la evaluación empírica de algoritmos de aprendizaje. Orange es fácil de usar para los inexpertos debido a su programación visual que se implementa a través de una interfaz en donde se trabaja con flujos mediante la vinculación de los widgets predefinidos o algunos creados por el propio usuario. Sin embargo, al ser una librería de Python, los usuarios más experimentados pueden valerse de la programación en ese lenguaje para realizar modificaciones más complejas.

Fue desarrollado originalmente por miembros de la Universidad de Liubliana, en Eslovenia, en el año 1996 como un framework para Aprendizaje de Máquina, escrito en C++. Con el tiempo se fueron creando añadidos para el framework desarrollado, módulos adicionales que fueron escritos en Python y con esto se denominó al framework como "Orange". (University of Ljubljana, 2019)

Tabla 2. Orange

Foco de Atención	Orientación al Aprendizaje de Máquina y Minería de Datos con principal interés en los flujos de trabajo que el usuario pueda definir.
Características principales	Simplicidad basada en la programación visual que está orientada a usuarios menos experimentados y permite una interacción rápida y efectiva con los datos
Lenguaje	Escrito en C, C++ y Python.
Interacción con el usuario	GUI para programación visual
Licenciamiento	GPLv3 o posteriores
Última versión estable	26 de octubre del 2018.

Principales características de la herramienta Orange (Elaboración Propia)

3.3.3 Motivos para la selección de Herramientas

Aunque existen muchas herramientas open source para análisis de datos, se han elegido a las dos citadas anteriormente debido, en primer lugar, al contraste que existe entre ambas. Debido a que se esperaría encontrar los mismos resultados con todas las herramientas, es decir, no importa dónde se genere un modelo, éste nos debería arrojar los mismos datos, se opta por seleccionar dos herramientas que trabajen de forma distinta para poder comparar el proceso. En el caso de R, éste lenguaje de programación se presta para la generación de gráficos, que es uno de sus pilares, así como el enfoque hacia la Ciencia de Datos y cuenta con una gran variedad de bibliotecas que se pueden usar de manera gratuita, luego, escoger el IDE RStudio se debió a la estabilidad del mismo y a las muchas opciones que presenta como tener múltiples ventanas dentro de la interfaz para poder trabajar en paralelo. En cuanto a Orange, ésta herramienta se escogió debido a que contrasta perfectamente con lo que ofrece R debido a que su funcionamiento se basa en programación visual, que no demanda conocimiento alguno de programación para poder generar modelos y está enfocada en la Minería de Datos. Otro motivo para escoger R es que Orange es en parte una biblioteca de Python, por lo que muchas otras herramientas basadas en este lenguaje fueron descartadas.

Debido a la tendencia que existe en la actualidad de llevar el trabajo del software hacia la nube y tener un modelo de Software as a Service, muchas personas se cuestionarían el hecho de usar ésta clase de herramientas en lugar de las clásicas aplicaciones para Windows o Linux. En el presente trabajo se justifica el hecho de usar software instalable en Windows debido a posibles problemas de conexión por la gran cantidad de datos con los que se está trabajando, que son una decena de archivos csv con muchas filas de datos en cada uno, lo que podría causar problemas a la hora de tener que conectar un cliente con los servidores de dichas aplicaciones y subir, cargar y descargar los datos correspondientes. Además, la mayoría de estos servicios en la nube son pagados, lo que van en contra de una de las condiciones con las que se empezó este trabajo (obtener herramientas open source y gratuitas).

4. Aplicación de Minería de Datos

4.1 Introducción a las Herramientas Elegidas.

4.1.1 Minería de Datos con R y RStudio

Existe una gran cantidad de trabajos previos en lo que se refiere a la aplicación de Minería de Datos usando el lenguaje de programación R y el IDE RStudio. Como en el Capítulo 3 ya se mencionaron las principales características de las herramientas elegidas, no se ahondará en este tema, sino más bien en detallar en líneas generales el acercamiento de otros autores.

Primero es importante señalar que muchas de las aplicaciones que se pueden encontrar fueron realizadas a modo de ejemplo de lo que la herramienta puede conseguir y para ello se usaron grupos de datos (datasets) predefinidos como el conocido dataset IRIS acerca de plantas y sus diferentes tipos que se emplea para problemas de Clasificación. También se usan otros grupos de datos populares. (Zhao, 2012)

Un estudio interesante en este tema lleva por nombre "*R and Data Mining: Examples and Cases Studies*" y es básicamente un manual de cómo se pueden ejecutar proyectos de Minería de Datos con esta herramienta. Lo interesante de este trabajo es que se estructura en base a los pasos propuestos para un proyecto de Minería de Datos, por lo que trata temas de preprocesamiento, visualización, etcétera, así como también muestra distintos casos como Clasificación, Análisis Predictivo, Clustering, Árboles de decisión, etcétera.

Empieza por una explicación breve de cómo se debe cargar la data en la herramienta, aspecto fundamental pues los datos son nuestro recurso principal y poder importarlos directamente desde un archivo .csv o .xlsx (Excel) indica que no serán necesarias conversiones de tipo y esto a la vez puede implicar que el trabajo de preprocesamiento no se alargue de más. Estos temas, así como explicaciones breves del manejo general de R y RStudio se tratan hasta el capítulo 2. El capítulo 3 se enfoca en mostrar las propias características de R para desplegar gráficos como un histograma y explorar los datos viendo, por ejemplo, cómo un grupo de datos se encuentra estructurado. Para el caso en el que se necesite el despliegue de gráficos en 3D se hace uso de una biblioteca que se añade al proyecto y que lleva por nombre: "scatterplot3d" así como algunas otras como "rgl", "lattice" y una biblioteca interesante que lleva por nombre "ggplot2" que sirve para desplegar gráficos complejos y exploración de datos.

Con respecto a la aplicación per sé de Minería de Datos, desde el capítulo 4 del mencionado trabajo se puede encontrar información relacionada. El trabajo realizado se puede resumir en la utilización de distintas bibliotecas que se ajusten a la problemática especificada. Por ejemplo, se empieza con la utilización de los algoritmos "Árboles de Decisión" y "Random Forest" para Análisis Predictivo y se basa el llamado de una función que genere el modelo y posteriormente una función que genere las predicciones. Por ejemplo, se usa la función "ctree()" de la biblioteca "party" para construir un árbol de decisión y luego se usa la función "predict()" para generar las predicciones correspondientes. Asimismo, se muestra cómo se particiona el conjunto de datos para poder trabajar (en problemas de Aprendizaje Supervisado); esto se aplica para todos los algoritmos tratados (variando la biblioteca y funciones usadas).

El tema del "scoring" de cada modelo se puede observar en los casos de estudio incluidos y se aprecia la inclusión de varias métricas con las cuales se puede dar un resultado a lo construido con el o los algoritmos elegidos. Precisamente los casos de estudio relacionados con predicciones son tratados por medio del algoritmo de Árboles de Decisión. (Zhao, 2012)

Analizando los casos de estudio del trabajo mencionado y toda la explicación previa en forma de pasos a seguir para alcanzar los objetivos en forma de modelos funcionales, se puede apreciar cómo la estructura de un proyecto de Minería de Datos es respetada y se parte desde el entendimiento mismo de los datos hasta generar una calificación o puntuación de la aplicación de un algoritmo apoyándose en el uso de bibliotecas adicionales para obtener funcionalidades extra que permitan cumplir de la manera más óptima con cada etapa del proceso.

4.1.2 Minería de Datos con Orange

Debido a la naturaleza de la herramienta Orange (considerándola como herramienta independiente enfocada en la programación visual y no como una biblioteca de Python) los textos basados en este software no son tan numerosos como con R, sin embargo, existen varios análisis como el del artículo denominado "Analysis of Data Using Data Mining tool Orange". Lo más destacado de este trabajo es la inclusión de una comparación entre varias herramientas de Minería de Datos como Weka, Knime, Data Melt o el mismo R con varios aspectos como si la plataforma es independiente, si es open source, si cuenta con una caja de herramientas amplia, etcétera. También muestra un poco del preprocesamiento de datos en base a una de las opciones de Orange para el despliegue de datos y elección de variables que participarán en el

correspondiente análisis. (Kukasvadiya & Divecha, Analysis of Data Using Data Mining tool, 2017)

Otros trabajos como el denominado "*Orange: Data Mining Fruitful and Fun - A Historical Perspective*" se centran en explicar los orígenes de la herramienta, así como sus principales enfoques como lo es la exploración de datos y la implementación de una interfaz sencilla de usar para todos los usuarios por medio de interacciones gráficas. Dadas las limitantes de la herramienta con un conjunto cerrado de algoritmos, no se suele elegir esta herramienta para cumplir tareas demandantes o asignadas a un negocio del mundo real, sin embargo, también cuenta con muchas opciones de visualización que hacen a la herramienta especialmente atractiva, como ya se mencionó, en la exploración de datos. Es considerablemente más sencillo usar una de las opciones de Orange para construir, por ejemplo, un histograma que hacerlo directamente por código como en la mayoría de herramientas como R. Por esto, Orange ha sido usado como un medio de testing para nuevas herramientas que se lancen al mercado; aunque, por supuesto, existen excepciones donde negocios tienen esta herramienta como principal soporte. (Demšar & Zupan, 2012)

El acercamiento real de la Minería de Datos con esta herramienta simplemente se basa en el uso de pequeños objetos visuales dentro de un "canvas" llamados widgets los cuales realizan una acción en concreto. Se cargan los datos, se los puede preprocesar, se los puede visualizar, se elige un algoritmo y se pueden obtener resultados en forma de métricas con las cuales analizar el proceso. Todo esto se puede constatar en la documentación oficial de Orange en su página web, la cual consta dentro de las referencias de este estudio.

4.1.3 Aproximación para el presente trabajo

Lo que se pretende realizar con este trabajo es similar a lo mencionado con el estudio analizado previamente con respecto a R y RStudio, es decir, se buscará crear modelos sencillos usando distintas bibliotecas que estén disponibles y posteriormente calificar dichos modelos en base a la obtención de distintas métricas. Se especificará paso a paso lo realizado para que cualquier persona pueda replicar el proceso y finalmente se compararán los resultados obtenidos. La idea general es crear modelos con diferentes casos que involucren distintas variables e ir mejorando los resultados con la inclusión o exclusión de algunas de esas variables. No se pretende crear modelos totalmente óptimos sino mas bien mostrar cómo se puede continuar optimizando los resultados.

Con respecto a Orange, debido a la facilidad que presenta esta herramienta, lo que se pretende es crear flujos de trabajo sencillos que cumplan con lo mismo que se realizó con la herramienta anterior, pero de manera gráfica. Debido a que existen varias maneras de desarrollar un modelo, se buscará aplicar una estructura que sea sencilla, entendible y además que de pie a la inclusión de más algoritmos en el mismo modelo para tener distintos resultados de métricas y ver cuál de los mencionados algoritmos se aplica mejor para el conjunto de datos usado.

4.1.4 Aclaración con respecto a las métricas

El uso de las métricas, detallado posteriormente, debe ser aclarado para evitar futuras confusiones. Las métricas presentadas en este y futuros capítulos no pretenden demostrar que una herramienta es mejor que otra. En todos los modelos a realizar se partirán de los mismos conjuntos de datos, se aplicará el mismo proceso y para la mayoría se usarán los mismos algoritmos. Se esperan, por lo tanto, resultados exactamente iguales entre ambas herramientas, con excepción de aquellos modelos en donde se apliquen otros algoritmos.

4.1.5 Objetivos y la generación de predicciones

El objetivo principal de esta aplicación es la de generar predicciones en los datos que fueron previamente seleccionados y analizados para posteriormente construir información que pueda ser de utilidad en un futuro, es decir, la aplicación se basarán en Análisis Predictivo. Como objetivo secundario se tiene el hecho de poder comparar dos herramientas que permiten realizar lo mismo (la minería), pero desde dos puntos de vista distintos y de esta forma poder emitir conclusiones acerca de con qué herramienta se obtienen mejores resultados. Y, asimismo, poner en práctica algunos de los métodos estudiados en el capítulo 2 del presente trabajo. Si bien es cierto que el conjunto de datos es amplio y que previamente se plantearon algunas preguntas que podían servir como ejemplo de los patrones derivados a encontrar, a continuación, se muestran definitivamente hacia lo que se orientará el análisis, de modo que las herramientas puedan iniciar una futura comprobación de dichas hipótesis, sin embargo, las hipótesis que se presentan a continuación no son parte de los objetivos de este trabajo, solamente se las incluyen para delimitar el trabajo y elaborar los modelos:

- El número de estudiantes mujeres que abandonan sus estudios tiene más presencia en las provincias de la Costa que en la Sierra.

- El número de estudiantes no promovidos (hombres) del octavo año de educación básica (primer año del "colegio" o secundaria) tiene relación con el sostenimiento del plantel educativo.
- El número de estudiantes varones promovidos del Segundo Año de Educación Básica tiene relación con el número de docentes mujeres y con el tipo de jornada en la que se encuentran estudiando.

4.2 Aplicación de Minería de Datos usando R

La primera aplicación de Minería de Datos se realizará con el lenguaje de programación R y la herramienta R Studio, elementos que se mencionaron en el capítulo previo. Se especificarán los pasos seguidos para obtener los resultados mediante capturas de pantalla y explicaciones cortas que permitan al lector reproducir el proceso con un conjunto de datos distinto.

4.2.1 Preprocesamiento

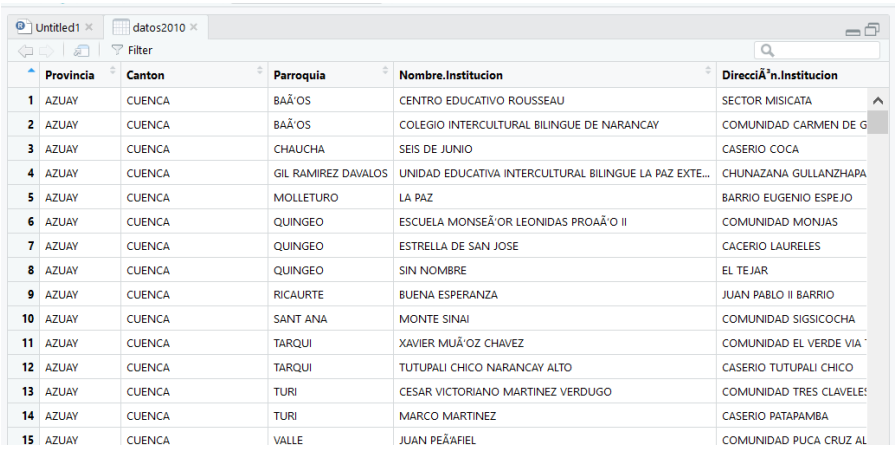
Como primer paso en la aplicación de Minería de Datos se debe hacer una limpieza de los datos por medio del preprocesamiento. Vale aclarar que ciertas acciones realizadas en el capítulo 3 del presente trabajo, sirven para acortar la duración de esta etapa. Por ejemplo, los datos ya fueron pasados por una revisión para eliminar las columnas que no eran de interés para el estudio, así como el análisis de si existían columnas con todos sus valores en cero, etcétera. Para empezar con el preprocesamiento en R Studio, se tiene que importar la data. El conjunto de datos debe estar en un archivo CSV para disminuir complejidad en el trabajo. La conversión de un archivo .xlsx (Excel) a un archivo CSV es sumamente sencilla y puede realizarse desde el mismo Excel.

```
> read.csv("C:/Users/sergi/Desktop/Tesis/Datos AMIE/2009-2010/Registros-Administrativos-2009-2010-
fin-AX.csv")->datos2010
```

Figura 17. Importación de datos en RStudio (Elaboración Propia)

En la Figura 17 se muestra la forma de leer un archivo CSV desde una ubicación específica en el equipo, así como la asignación de un nombre para seguir trabajando con dicho archivo. El archivo que se muestra corresponde a los datos del período lectivo 2009-2010. Una vez se tienen los datos importados en R Studio, se procede a mirarlos con el fin de tener una

mejor gestión de los mismos, esto se realiza mediante la función “View” y se obtiene algo similar a lo que se muestra a continuación:



Untitled1 | datos2010 | Filter

	Provincia	Canton	Parroquia	Nombre.Institucion	DirecciÃ³n.Institucion
1	AZUAY	CUENCA	BAÃ´OS	CENTRO EDUCATIVO ROUSSEAU	SECTOR MISICATA
2	AZUAY	CUENCA	BAÃ´OS	COLEGIO INTERCULTURAL BILINGUE DE NARANCAY	COMUNIDAD CARMEN DE G
3	AZUAY	CUENCA	CHAUCHA	SEIS DE JUNIO	CASERIO COCA
4	AZUAY	CUENCA	GIL RAMIREZ DAVALOS	UNIDAD EDUCATIVA INTERCULTURAL BILINGUE LA PAZ EXTE...	CHUNAZANA GULLANZHAPA
5	AZUAY	CUENCA	MOLLETURO	LA PAZ	BARRIO EUGENIO ESPEJO
6	AZUAY	CUENCA	QUINGEO	ESCUELA MONSEÃ´OR LEONIDAS PROAÃ´O II	COMUNIDAD MONJAS
7	AZUAY	CUENCA	QUINGEO	ESTRELLA DE SAN JOSE	CACERIO LAURELES
8	AZUAY	CUENCA	QUINGEO	SIN NOMBRE	EL TEJAR
9	AZUAY	CUENCA	RICAURTE	BUENA ESPERANZA	JUAN PABLO II BARRIO
10	AZUAY	CUENCA	SANT ANA	MONTE SINAI	COMUNIDAD SIGSICOCHA
11	AZUAY	CUENCA	TARQUI	XAVIER MUÃ´OZ CHAVEZ	COMUNIDAD EL VERDE VIA
12	AZUAY	CUENCA	TARQUI	TUTUPALI CHICO NARANCAY ALTO	CASERIO TUTUPALI CHICO
13	AZUAY	CUENCA	TURI	CESAR VICTORIANO MARTINEZ VERDUGO	COMUNIDAD TRES CLAVELES
14	AZUAY	CUENCA	TURI	MARCO MARTINEZ	CASERIO PATAPAMBA
15	AZUAY	CUENCA	VALLE	JUAN PEÃ±AFIEL	COMUNIDAD PUCA CRUZ AL

Figura 18. Visualización de Datos Importados en RStudio (Elaboración Propia)

Una vez que se tienen los datos importados, será necesario importar la biblioteca “dplyr” que sirve para la manipulación de datos y con la cual podremos gestionar el archivo fuente de manera simple.

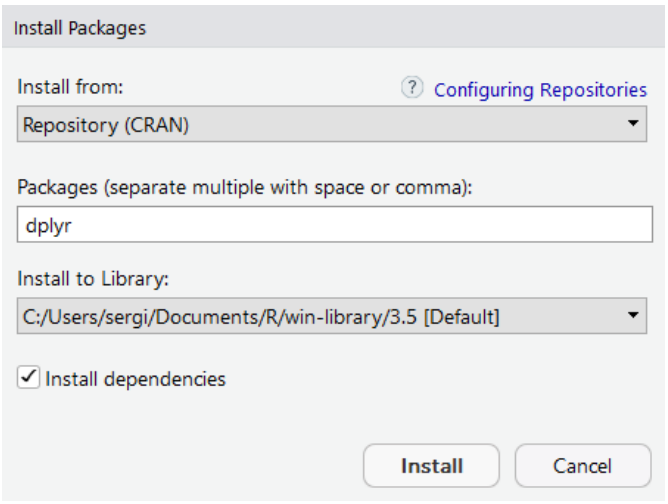


Figura 19. Ventana de instalación de paquetes adicionales en RStudio (Elaboración Propia)

Luego de que la biblioteca se descargue en el equipo, será necesario importarla mediante el comando “library” y se obtendrá un mensaje como el que se muestra en la Figura 20.

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Figura 20. Importación de la biblioteca "dplyr" (Elaboración Propia)

Lo que usaremos de la biblioteca dplyr será la selección de datos para no tomar en cuenta algunas columnas. A pesar de que previamente ya se eliminaron columnas sin importancia para el análisis, aún permanecen algunas que terminaron por ser irrelevantes con la definición de los patrones a buscar. Para ello, se tendrá que introducir el siguiente comando:

```
> datos2010 %>% select(c(-1))->datos2010
> datos2010 %>% select(c(-5,-6))->datos2010
> datos2010 %>% select(c(-5))->datos2010
```

Figura 21. Selección de datos usando la biblioteca "dplyr" (Elaboración Propia)

Mediante "select" y especificando el número de columna como parámetro se puede decidir qué columnas no serán usadas y grabarlas sobre la misma variable "datos2010".

El preprocesamiento de este archivo será terminado en este punto debido a que los datos ya están estructurados, consistentes y claros para seguirlos analizando. Se podría seguir manipulando la data con la biblioteca usada o con alguna otra similar, de hecho, varios de los ajustes que se hicieron manualmente desde Excel bien podrían hacerse en R Studio de manera directa, esto ahorraría algo de tiempo.

4.2.2 Visualización de los Datos

Para visualizar los datos se hará uso de la biblioteca "ggplot2", la cual se descargará gratuitamente de la misma manera que la biblioteca "dplyr", así como su importación al proyecto. Este paquete fue creado por Hadley Wickham y consiste en un lenguaje gráfico para mostrar datos de una manera elegante; su desarrollo se basó en The Grammar of Graphics de Leland Wilkinson. (Kabacoff, 2017)

Con esta biblioteca se añaden funciones interesantes que sirven para mostrar los datos mediante gráficos útiles. En primer lugar, se debe tomar en cuenta que muchas de estas

bibliotecas añadidas de R, si bien conservan la lógica general del lenguaje, manejan sus propias reglas internas, por lo que es menester de cada programador el averiguar su funcionamiento. Para demostrar cómo se comportan los datos seleccionados para este trabajo, la función de esta biblioteca que usaremos como principal herramienta será "ggplot" que es precisamente la que permite desplegar los datos. A continuación, en la Figura 22, se muestra un ejemplo de esto usando dos columnas del conjunto de datos con el que se está trabajando:

```
datos1 <- ggplot(data=datos2010,aex(y=HombresPromovidosTerceroBach,x=Regimen.Escolar))+geom_poin
t()
```

Figura 22. Usando la biblioteca "ggplot2" (Elaboración Propia)

En primer lugar, se muestra cómo se asigna una determinada operación a una variable, que en este caso representa un nuevo conjunto de datos llamados "datos1". La operación que se está llevando a cabo es la de formar un gráfico con ggplot. Esta función toma como parámetros:

- Un conjunto de datos (data =datos2010). Se indica que los datos fuente provienen de nuestro archivo principal que pasó por el preprocesamiento en el punto anterior.
- Especificación de los ejes cartesianos del gráfico, en este caso se puede ver claramente cómo la columna de los estudiantes varones promovidos del tercer año de Bachillerato constan como el eje y, mientras que el Régimen Escolar consta como el eje x.
- El tipo de gráfico que se desplegará. El tipo de gráfico geom_point() se usa como ejemplo en este caso.

Empezando con un ejemplo más sencillo donde solo se tome en cuenta una columna, se puede construir un histograma, sin embargo, antes de realizar esto es conveniente señalar que el programador puede controlar los límites que se muestran en el gráfico, esto con el fin de obtener una representación correcta. Para esto simplemente se escribe: "$datos1 + ylim(0{,}800)$" en donde "lim" establece los límites superior e inferior, en este caso del eje y, que irían desde el 0 hasta el 800.

```
datos1 <- ggplot(data=datos2010,aes(x=Docentes.Femenino))+geom_histogram(bins=40)
datos1 + ylim(0,500) + xlim(0,250)
```

Figura 23. Generación de un histograma del número de docentes mujeres usando la biblioteca "ggplot2" en RStudio. (Elaboración Propia)

La Figura 23 muestra cómo se establece el comando para generar el histograma (en la primera línea), para posteriormente limitar el gráfico de acuerdo al criterio del programador, con el fin de obtener algo entendible.

```
Warning messages:
1: Removed 1 rows containing non-finite values (stat_bin).
2: Removed 5 rows containing missing values (geom_bar).
```

Figura 24. Mensajes de advertencia en RStudio al usar la biblioteca "ggplot2" (Elaboración Propia)

Existen algunos casos en los que se encuentran valores no finitos entre la colección de datos usada para generar el gráfico, esto lleva a eliminar dichos valores y recibir un mensaje de precaución. De igual manera, se recibe esta clase de mensajes cuando los límites establecidos no contienen todos los valores del conjunto de datos.

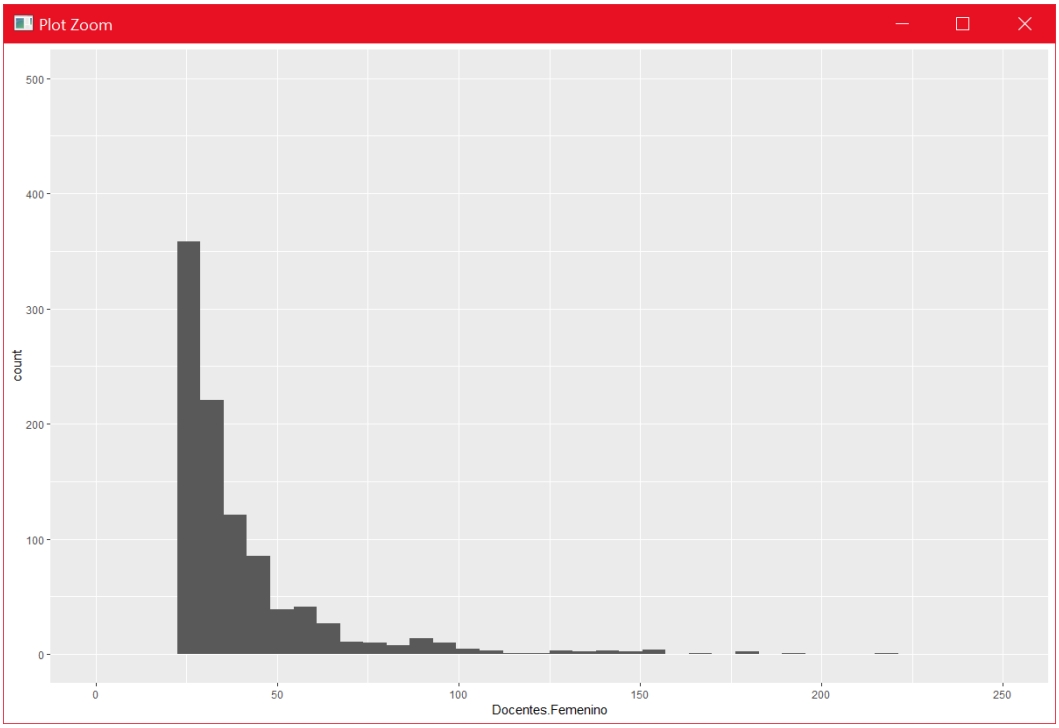


Figura 25. Histograma con datos de docentes mujeres obtenido mediante la biblioteca "ggplot2" en RStudio (Elaboración Propia)

La Figura 25 muestra el resultado esperado, un histograma. En el eje x se muestran los números de los docentes mujeres en las instituciones educativas del país. El histograma nos indica que en la mayoría de instituciones, el número de mujeres no supera las 25 personas, mirando sin mucho detalle al gráfico, se podría decir que el número de 25 docentes mujeres se repite en poco más de 350 instituciones y que muy pocas instituciones tienen más de 150

docentes mujeres en sus planteles. Evidentemente ejecutar un comando para obtener un gráfico se presenta como simple y poco explicativo. Por suerte, esta biblioteca nos deja combinar tipos de gráficos mediante el operador "+". En la Figura 26, se muestra la relación entre los estudiantes varones promovidos del segundo año de educación básica y el número de docentes femeninos, indicando que existe son proporcionales, es decir, a más estudiantes varones promovidos del segundo año de básica, más docentes mujeres existen. El gráfico se logra mediante la combinación de geom_points() y geom_smooth().

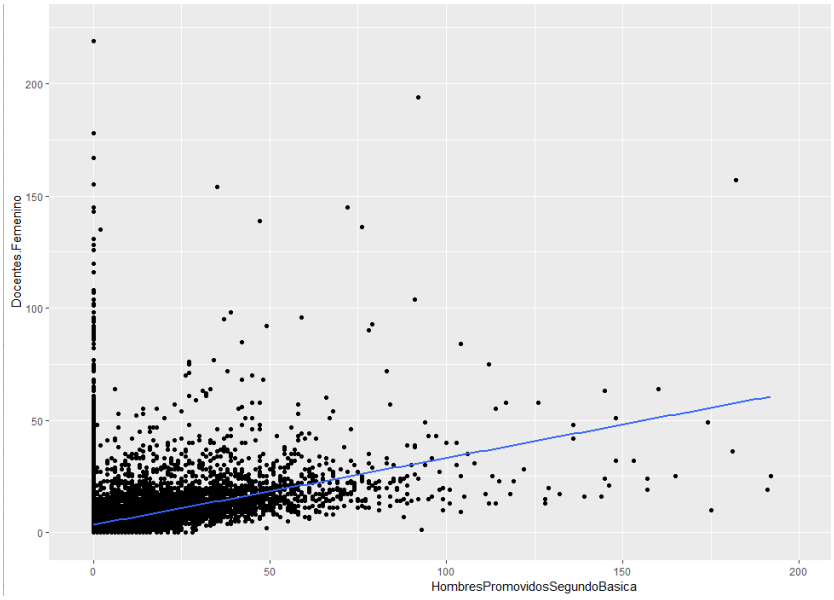


Figura 26. Relación entre los Docentes Femeninos y los estudiantes varones promovidos del segundo año de educación básica (Elaboración Propia)

De la misma manera que se pueden corroborar hipótesis previas, también se pueden descartar algunas ideas adicionales que se desprenden de esas mismas hipótesis. Por ejemplo, se podría pensar que existe alguna clase de relación entre los administrativos mujeres y el gráfico anteriormente mostrado en la Figura 26. Para llevar a cabo esto, se plantea lo mostrado en la Figura 27:

```
> datos1 <- ggplot(data=datos2010,aes(x=HombresPromovidosSegundoBasica, y=Docentes.Femenino,col=fa
ctor(Administrativos.Femenino)))+geom_point()+geom_smooth(method="lm",se=F)+labs(col="Administrati
vos.Femenino")
> datos1 + ylim(0,250) + xlim(0,250)
```

Figura 27. Gráfico que relaciona las variables de la Figura 26 más la variable de los Administrativos Mujeres. (Elaboración Propia)

Como se puede apreciar, los ejes en el gráfico son exactamente los mismos, para mantener coherencia, sin embargo, se añade la nueva variable indicada. También se puede notar cómo se puede personalizar los gráficos y que se mantienen los límites previamente mencionados.

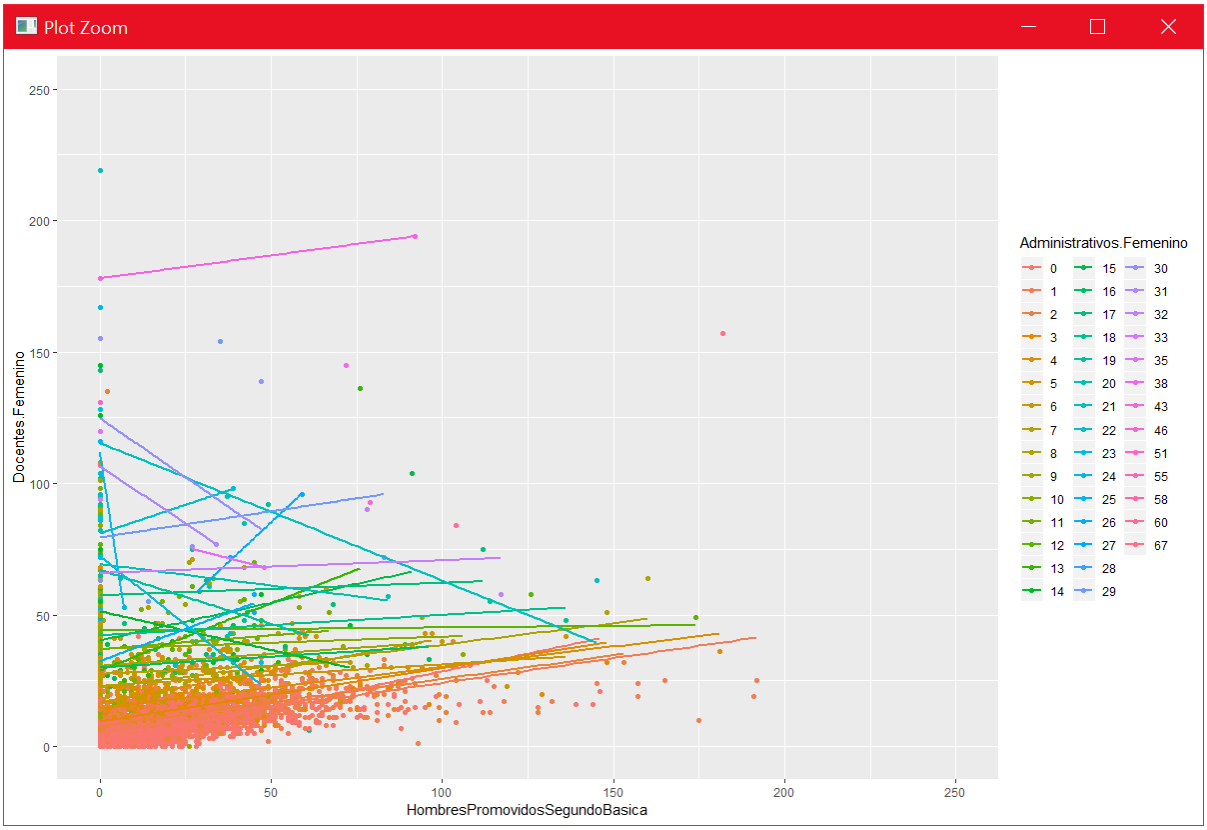


Figura 28. Gráfico resultante del código descrito en la Figura 27. (Elaboración Propia)

La Figura 28 nos presenta los resultados y de inmediato se entiende que el número de administrativos femenino varía muchísimo de un caso a otro, indicando que no existe repercusión de esta variable con los estudiantes varones promovidos del segundo año de básica. La leyenda en la parte derecha nos muestra los números de los administrativos mujeres en forma de distintas rectas que pueblan el gráfico, en donde existe gran diversidad.

4.2.3 Minería de Datos (Modelado y Evaluación de Datos)

Para la aplicación de Minería de Datos se usará el método de Regresión Lineal, que fue detallado en el capítulo 2 del presente trabajo. Como se mencionó en ese capítulo, la Regresión Lineal es un método de Aprendizaje Supervisado y éste se caracteriza por requerir datos de entrenamiento para llevar a cabo el aprendizaje. Por esta razón es que se requiere dividir los datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de pruebas, que a la postre también nos servirá para probar el modelo a desarrollar.

Para empezar, se hará uso de la biblioteca “caTools” la cual se debe descargar y agregar al proyecto de la misma manera que con las bibliotecas usadas previamente. Justamente la separación de los datos en los dos subconjuntos es el primer paso. La función split de la

biblioteca que se acaba de añadir es la que permite ejecutar esta acción. La forma en que esto sucede es que los datos se dividen en base a dos etiquetas: TRUE y FALSE. Se debe establecer qué porcentaje de los datos fuente tendrán la etiqueta TRUE. Esto se ejemplifica en el código de la Figura 29:

Figura 29. Función para separar los datos en base a una variable aleatoriamente elegida (Elaboración Propia)

El código mostrado corresponde a la función "sample.split" la cual requiere particionar los datos en base a una variable de este conjunto, la cual viene a ser una de las columnas del archivo original, en este caso "HombresPromovidosTerceroBach" que es la cantidad de hombres promovidos del Tercero de Bachillerato. El siguiente parámetro es el "SplitRatio" que, como se dijo previamente, es el porcentaje que se indica para dividir los datos en dos conjuntos en donde el sesenta y cinco por ciento de los datos tendrán etiqueta TRUE y el resto FALSE; luego, se almacena en una variable llamada "split_index".

Ahor bien, los datos ya están divididos virtualmente, pero se necesita una separación real. Para lo cual se procede a crear dos variables y en ellas almacenar todos los datos correspondientes. En el primer conjunto "train" (por "entrenamiento") será llenado por aquellos datos con la etiqueta TRUE y el conjunto "test" (por "prueba") con los datos con la etiqueta FALSE. Posteriormente procedemos a verificar el número de filas que se tiene en cada subconjunto, tal y como se muestra en la Figura 30.

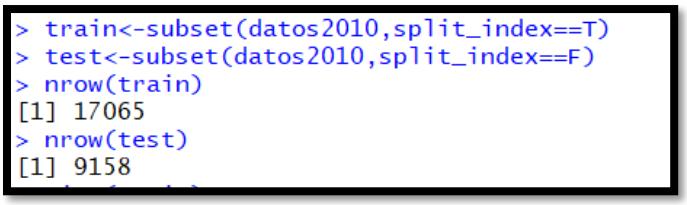


Figura 30. Separación de los datos en dos subconjuntos "train" y "test" mediante las etiquetas de TRUE o FALSE establecidas previamente. (Elaboración Propia)

Estos se convierten en los conjuntos de datos que se procederán a usar. Una vez se tienen los datos divididos, se procede con la construcción del modelo de Regresión Lineal. Para comenzar, tenemos que definir una variable que sea dependiente, que en este caso será aquella que sea el centro del análisis. En este caso se establecerá como variable dependiente a los Hombres Promovidos del Segundo Año de Educación Básica por ser una variable que

técnicamente está con valores distintos a cero en todas las filas de datos (ya que todas las instituciones tienen primaria) y sirve como parte de una de la resolución de una de las hipótesis planteadas.

Figura 31. Aplicación de la función lm con la variable HombresPromovidosSegundoBasica como variable dependiente. (Elaboración Propia)

La Figura 31 muestra cómo se aplica la función "lm" que viene a ser un modelo lineal, que es en lo que se basa la Regresión Lineal. Para aplicar esta función se debe tener una variable objetivo que es dependiente del resto. Para indicar que todo el resto de variables ha de ser considerada en el modelo se usan los símbolos "c" y a continuación se especifica el conjunto de datos a considerar, que en este caso es el conjunto de entrenamiento. Todo este proceso se guarda en la variable "mod1" que representa el modelo final. Una vez se tiene el modelo generado, se procede a predecir los valores. El tema de predecir valores es fundamental y está estrechamente ligado con el Aprendizaje de Máquina. La máquina recibió los valores de entrenamiento que pasan por un modelo de Regresión Lineal para poder predecir nuevos valores. Esto se realiza tal y como se muestra en la Figura 32.

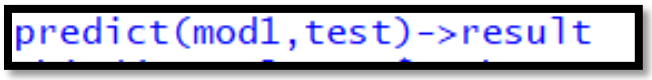


Figura 32. Predicción de nuevos valores usando la función "predict" en RStudio. (Elaboración Propia)

Con la función "predict" se realiza lo antes mencionado, es decir, la predicción de nuevos valores. Para lograrlo, se deben usar los argumentos de un modelo (que ya se generó) y un nuevo conjunto de datos, que es el conjunto de datos de pruebas. Esto se guarda en una nueva variable llamada "result". Esta variable result se vuelve un nuevo conjunto de datos que almacena todos las predicciones realizadas. Para poder tener una idea más gráfica de los resultados, se combinan los valores actuales y los que son resultado de las predicciones en una misma tabla y se los muestra en pantalla, esto mediante lo indicado en la Figura 33.

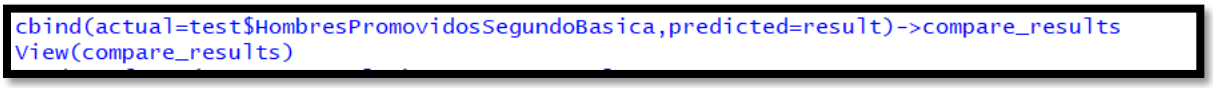


Figura 33. Comparación de los resultados actuales y los predichos de los datos obtenidos en RStudio. (Elaboración Propia)

Básicamente, la función "cbind" permite la conjunción de dos conjuntos de datos. En este caso, se está escogiendo a la columna de los Hombres Promovidos del Segundo Año de Básica debido a que ésta fue la que se tomó como variable dependiente, y se los junta con los datos predichos provenientes de la variable "result", para posteriormente guardarlos en la variable "compare_results".

	actual	predicted
1	2	4.52109484
4	0	-0.58663825
7	0	0.88560872
8	3	4.11790277
9	10	6.56372195
11	4	3.62654764
12	0	0.05898625
13	5	5.22568467
15	0	2.27229376
17	3	2.91583975
18	0	4.70023666
20	0	4.80305840
21	3	4.33518555
34	0	0.87189674

Figura 34. Resumen de la tabla de comparación de resultados (actuales y predichos) de lo obtenido en RStudio. *(Elaboración Propia)*

La Figura 34 muestra la comparación esperada. Sin embargo, para tener una idea más clara de qué tan acertada es la predicción, se procede a calcular el error entre cada par de valores. Antes, se debe hacer una transformación de la variable "compare_results" para que pase a ser un data frame para poder aplicar la debida operación que viene a ser una simple resta entre el valor actual menos el valor predicho. Se vuelven a juntar los datos para tener la comparación gráficamente y se muestra por pantalla, tal y como se indica en la Figura 35.

```
as.data.frame(compare_results)->compare_results
compare_results$actual-compare_results$predicted->error
cbind(compare_results,error)->resultados_a_comparar
View(resultados_a_comparar)
```

Figura 35. Procedimiento en forma de comandos para mostrar el error de los datos actuales y los predichos en RStudio. (Elaboración Propia)

El error obtenido en cada uno de los pares de datos se muestra en la Figura 36, indicando que no se presentan errores considerables, cosa que será corroborada en la evaluación posterior del modelo realizado.

	actual	predicted	error
1	2	4.52109484	-2.52109484
4	0	-0.58663825	0.58663825
7	0	0.88560872	-0.88560872
8	3	4.11790277	-1.11790277
9	10	6.56372195	3.43627805
11	4	3.62654764	0.37345236
12	0	0.05898625	-0.05898625
13	5	5.22568467	-0.22568467
15	0	2.27229376	-2.27229376
17	3	2.91583975	0.08416025
18	0	4.70023666	-4.70023666
20	0	4.80305840	-4.80305840
21	3	4.33518555	-1.33518555
34	0	0.87189674	-0.87189674

Figura 36. Resumen de la tabla de comparación de resultados con su error respectivo del trabajo realizado en RStudio. (Elaboración Propia)

Para tener una idea más aproximada del error obtenido, se hace uso del Error Cuadrático Medio, que también nos dice la diferencia entre algún valor estimado y valores reales. Para ello se emplea la función "sqrt" y "mean" en base al data frame "compare_results" y su variable "error", esto se guarda en la variable "error_cuadratico_medio1", tal y como se muestra en la Figura 37.

```
sqrt(mean(compare_results$error^2))->errr_cuadratico_medio1
```

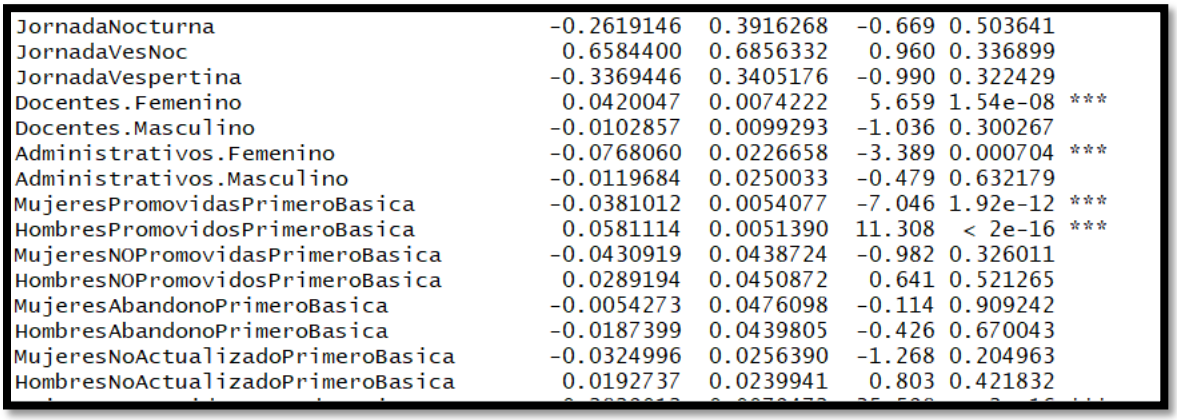
Figura 37. Obteniendo el error cuadrático medio en base a los valores obtenidos previamente en el trabajo con RStudio. (Elaboración Propia)

Luego de haber comprendido las salidas del modelo, que en este caso resultan en predicciones, se procede a realizar un análisis más a profundidad. Esto se logra usando una función llamada "summary" que provee un resumen del modelo, tal y como se muestra en la Figura 38.

```
> summary(mod1)

Call:
lm(formula = HombresPromovidosSegundoBasica ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-39.349  -1.210  -0.209   1.026  93.276

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.1682229  0.5071331   0.332 0.740110
```

Figura 38. Resumen del modelo realizado en RStudio. (Elaboración Propia)

Una de las utilidades de analizar estos resúmenes es que proveen información de qué tanto afecta una variable a otra para arrojar los datos predichos. Esto evidentemente nos ayuda para poder refinar el modelo, ya que, durante la llamada de la función del modelo, se pueden descartar variables. La forma de ver el nivel de influencia de una variable es simple, mostrándose en forma de estrellas en la parte derecha de cada variable. Si una variable tiene tres estrellas significa que tiene gran influencia para la predicción de datos. Esto se aprecia en la Figura 39.



JornadaNocturna	-0.2619146	0.3916268	-0.669	0.503641	
JornadaVesNoc	0.6584400	0.6856332	0.960	0.336899	
JornadaVespertina	-0.3369446	0.3405176	-0.990	0.322429	
Docentes.Femenino	0.0420047	0.0074222	5.659	1.54e-08	***
Docentes.Masculino	-0.0102857	0.0099293	-1.036	0.300267	
Administrativos.Femenino	-0.0768060	0.0226658	-3.389	0.000704	***
Administrativos.Masculino	-0.0119684	0.0250033	-0.479	0.632179	
MujeresPromovidasPrimeroBasica	-0.0381012	0.0054077	-7.046	1.92e-12	***
HombresPromovidosPrimeroBasica	0.0581114	0.0051390	11.308	< 2e-16	***
MujeresNOPromovidasPrimeroBasica	-0.0430919	0.0438724	-0.982	0.326011	
HombresNOPromovidosPrimeroBasica	0.0289194	0.0450872	0.641	0.521265	
MujeresAbandonoPrimeroBasica	-0.0054273	0.0476098	-0.114	0.909242	
HombresAbandonoPrimeroBasica	-0.0187399	0.0439805	-0.426	0.670043	
MujeresNoActualizadoPrimeroBasica	-0.0324996	0.0256390	-1.268	0.204963	
HombresNoActualizadoPrimeroBasica	0.0192737	0.0239941	0.803	0.421832	

Figura 39. Influencia de variables en el modelo realizado en RStudio. (Elaboración Propia)

Posteriormente se analizarán los datos obtenidos, sin embargo, se puede decir que existe lógica con lo obtenido. Existe una influencia directa entre el número de docentes mujeres y los estudiantes promovidos del primero de básica.

Aun obteniendo errores relativamente bajos y observando que la influencia de las variables mantiene una lógica con lo esperado, es necesario estar matemáticamente seguro de que el modelo funciona. Es por ello por lo que existe un valor llamado “R cuadrado ajustado” que es, básicamente una medida de bondad que sirva para identificar la precisión de modelos lineales. Si bien es cierto que este número tiende a aumentar a medida que existan más variables a considerar para generar las predicciones, analizar el valor obtenido sirve para tener una idea de si el modelo es o no útil. El caso óptimo es el del modelo con R Cuadrado Ajustado con valor igual a 1. El valor obtenido de 92,53% se considerable aceptable, sin embargo, se pueden eliminar las variables que en el resumen aparecen como poco influyentes para las predicciones deseadas y ver si el R Cuadrado Ajustado aumenta o disminuye. Estos resultados se observan en la Figura 40.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.452 on 16920 degrees of freedom
Multiple R-squared:  0.926,     Adjusted R-squared:  0.9253
F-statistic:  1469 on 144 and 16920 DF,  p-value: < 2.2e-16
```

Figura 40. Métricas correspondientes al modelo realizado en RStudio, entre ellos, el “R Cuadrado Ajustado”. (Elaboración Propia)

El siguiente paso por ejecutar fue el de probar eliminar algunas de las variables que en el resumen del modelo fueron catalogadas como sin impacto o con poco impacto para las predicciones. Se crearon algunos modelos basados en esto y se obtuvieron dos modelos con valores que determinan la funcionalidad de este método. En primer lugar, se generó un modelo “mod5” eliminando todas las variables que aparezcan en el resumen con dos, una o ninguna estrellas, pero respetando las variables que se cruzaban con la hipótesis relacionada (en este caso Jornada y Docentes). Teniendo en cuenta que, como se dijo previamente, a mayor número de variables más alto es el valor del R Cuadrado Ajustado, se puede decir que lo obtenido es considerable pues este modelo cuenta con un valor de 92.49% muy cercano al original. El siguiente modelo “mod6” se realizó eliminando todas las variables originales exceptuando las que se cruzaban con la hipótesis relacionada (de nuevo, Jornada y Docentes).

Los resultados se aprecian en la Figura 41, sin embargo, de entrada, se puede decir que no son óptimos, pues el R Cuadrado Ajustado tiene un valor muy bajo y poco aceptable para decir que el modelo predice valores correctamente. En los casos anteriores, se tenía que el modelo predecía correctamente 9 de cada 10 valores, en este caso se aprecia una caída a poco menos de 2 de cada 10 valores. Se podría pensar que el método de Regresión Lineal no se ajusta a este grupo de datos, sin embargo, esto no es completamente cierto ya que, como se dijo, los modelos previos arrojaron predicciones satisfactorias. Otra de las presunciones que podrían darse es que realmente las variables independientes de la hipótesis no tienen repercusión en la variable dependiente, o en otras palabras, el tipo de Jornada y el número de Docentes Mujeres no afecta al número de Hombres Promovidos del Segundo Año de Básica. Si bien es cierto que esto, en conjunto con los demás datos serán analizados en el siguiente capítulo, se puede adelantar, ya teniendo a la vista varios resultados de modelos con Regresión Lineal que, al menos la variable de Docentes Mujeres, sí tiene un impacto, debido a que en los resúmenes siempre aparece como de alta influencia.

Lo que nos deja como explicación el hecho de que, en realidad, las demás variables que se descartaron para el último modelo sí son influyentes y el modelo no puede construirse sin éstas.

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.53 on 17056 degrees of freedom
Multiple R-squared:  0.1675,    Adjusted R-squared:  0.1671
F-statistic:   429 on 8 and 17056 DF,  p-value: < 2.2e-16
```

Figura 41. Resultados de una nueva versión del modelo realizado en RStudio reduciendo las variables solamente a las relacionadas con la hipótesis. (Elaboración Propia)

Para el caso del abandono de mujeres del tercero de Bachillerato, se procede a trabajar de la misma manera que con el caso anterior; en este caso la variable dependiente será “MujeresAbandonoTerceroBach” y el resto de columnas en el grupo de datos original fungirán como variables independientes, sin exclusiones.

El resultado de lo obtenido se muestra en la Figura 42 y se puede apreciar un efecto con un valor de R Cuadrado Ajustado un poco bajo comparado con el caso anterior, sin embargo, también es aceptable.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.361 on 16920 degrees of freedom
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.7077
F-statistic: 287.9 on 144 and 16920 DF,  p-value: < 2.2e-16
```

Figura 42. Resultados del Primer Modelo en RStudio con la variable “MujeresAbandonoTerceroBach” como dependiente. (Elaboración Propia)

Para el segundo modelo se eliminarán todas las variables correspondientes a los estudiantes de primaria, ya que en el primer modelo se muestran como poco o nada influyentes. Los resultados, sin embargo, no indican una mejora en el modelo. Sin embargo, arrojan datos interesantes: no existe una influencia entre el número de mujeres que abandonan y la provincia a la que pertenece la institución educativa.

Para corroborar esto, se generará un tercer modelo y en el caso en que el RCA se mantenga o se eleve, quiere decir que la variable de Provincia no tiene un real impacto en la predicción de datos de la variable dependiente establecida. El resumen de este modelo se puede apreciar en la Figura 43. En base a lo obtenido, se puede decir que para mejorar el RCA se deberían explorar otros métodos como Clasificación. Esto debido a que, como se mencionó

anteriormente, los datos por su distribución no se ajustan perfectamente a todos los métodos disponibles para la Minería de Datos (y en general para el Aprendizaje de Máquina). Aunque vale recalcar una vez más, que los resultados del modelo generado no son despreciables.

```
HombresAbandonoNovenoBasica          0.0154969  0.0029175   5.312 1.10e-07 ***
MujeresNoActualizadoNovenoBasica    -0.0081048  0.0027356  -2.963 0.003054 **
HombresNoActualizadoNovenoBasica     0.0028834  0.0023561   1.224 0.221044
MujeresPromovidasDecimoBasica        0.0058129  0.0007390   7.866 3.89e-15 ***
HombresPromovidosDecimoBasica       -0.0007379  0.0007348  -1.004 0.315300
MujeresNOPromovidasDecimoBasica      0.0045116  0.0028370   1.590 0.111795
HombresNOPromovidosDecimoBasica     -0.0003834  0.0018283  -0.210 0.833907
MujeresAbandonoDecimoBasica          0.0819240  0.0043268  18.934  < 2e-16 ***
HombresAbandonoDecimoBasica         -0.0148135  0.0033318  -4.446 8.80e-06 ***
MujeresNoActualizadoDecimoBasica     0.0016514  0.0013572   1.217 0.223685
HombresNoActualizadoDecimoBasica    -0.0010994  0.0016502  -0.666 0.505299
MujeresPromovidasPrimeroBach        -0.0060492  0.0007363  -8.216 2.25e-16 ***
HombresPromovidosPrimeroBach         0.0058161  0.0007851   7.408 1.34e-13 ***
MujeresNOPromovidasPrimeroBach      -0.0087183  0.0019244  -4.530 5.93e-06 ***
HombresNOPromovidosPrimeroBach      -0.0007462  0.0016583  -0.450 0.652716
MujeresAbandonoPrimeroBach           0.0683132  0.0030205  22.616  < 2e-16 ***
HombresAbandonoPrimeroBach           0.0058499  0.0025229   2.319 0.020424 *
MujeresNoActualizadoPrimeroBach      0.0212039  0.0022287   9.514  < 2e-16 ***
HombresNoActualizadoPrimeroBach     -0.0037912  0.0014226  -2.665 0.007708 **
MujeresPromovidasSegundoBach        -0.0030014  0.0009232  -3.251 0.001152 **
HombresPromovidosSegundoBach        -0.0047084  0.0008602  -5.474 4.47e-08 ***
MujeresNOPromovidasSegundoBach      -0.0172376  0.0031020  -5.557 2.79e-08 ***
HombresNOPromovidosSegundoBach       0.0161080  0.0023461   6.866 6.84e-12 ***
MujeresAbandonoSegundoBach           0.1267092  0.0041863  30.267  < 2e-16 ***
HombresAbandonoSegundoBach          -0.0460817  0.0039382 -11.701  < 2e-16 ***
MujeresNoActualizadoSegundoBach     -0.0083077  0.0028044  -2.962 0.003056 **
HombresNoActualizadoSegundoBach      0.0009403  0.0023194   0.405 0.685198
MujeresPromovidasTerceroBach         0.0139871  0.0007816  17.895  < 2e-16 ***
HombresPromovidosTerceroBach        -0.0048233  0.0006509  -7.410 1.32e-13 ***
MujeresNOPromovidasTerceroBach       0.0202016  0.0066503   3.038 0.002387 **
HombresNOPromovidosTerceroBach      -0.0043451  0.0021864  -1.987 0.046900 *
HombresAbandonoTerceroBach           0.3074037  0.0060095  51.153  < 2e-16 ***
MujeresNoActualizadoTerceroBach     -0.0101921  0.0012570  -8.108 5.47e-16 ***
HombresNoActualizadoTerceroBach     -0.0004331  0.0005545  -0.781 0.434759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3609 on 16996 degrees of freedom
Multiple R-squared:  0.709,     Adjusted R-squared:  0.7078
F-statistic: 608.9 on 68 and 16996 DF,  p-value: < 2.2e-16
```

Figura 43. Modelo en RStudio actualizado con la variable Mujeres Abandono Tercero de Bachillerato donde se mantiene el valor del R Cuadrado Ajustado. (Elaboración Propia)

Por último, se genera un modelo para la hipótesis relacionada con lo estudiantes varones del Octavo de Básica. Para esto, se repite el proceso que en los dos casos anteriores: se genera un modelo solamente especificando la variable dependiente y luego se analiza qué variables eliminar para seguir puliendo el modelo. El primer modelo obtenido cuenta con un R Cuadrado Ajustado de 83.26% que es un valor aceptable y se puede decir que el modelo funciona. Lo que se puede obtener de un vistazo rápido del resumen del modelo es que las variables de los estudiantes (hombres y mujeres) de primaria no tienen un impacto en las predicciones. Lo que se procede a realizar es eliminar del modelo a estas variables y lo que se obtiene es un modelo con un valor de RCA de 83.28% que representa un valor casi igual que el anterior, con lo que se confirma que las variables eliminadas realmente no tenían impacto. Estos valores se muestran en

la Figura 44 en donde se aprecia el resumen del segundo modelo con respecto a las variables de la hipótesis relacionada.

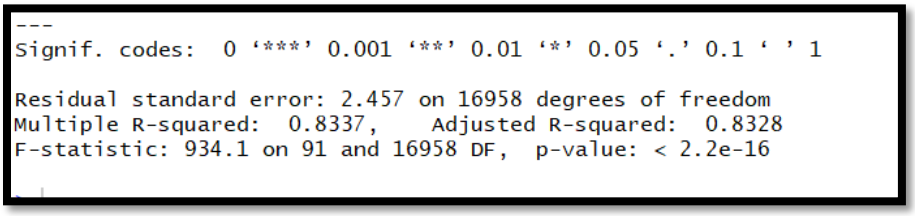


Figura 44. Resumen del segundo modelo aplicado con variable dependiente como "HombresNOAprovadosOctavoBásica" y eliminando las variables que en el primer modelo se consideraron como poco influyentes. (Elaboración Propia)

Una vez se tienen los modelos necesarios para tratar de construir una vía de respuesta a las tres hipótesis planteadas, se da por terminado el proceso con R y R Studio. Este proceso debe ser replicado para cada uno de los demás archivos involucrados en este análisis con el fin de recopilar los datos necesarios para compararlos en el siguiente capítulo y obtener las conclusiones respectivas.

4.3 Aplicación de Minería de Datos usando Orange

La aplicación con Orange es un tanto más directa y visual, especialmente por el hecho de que Orange es una herramienta de programación visual y no requiere de la escritura de código por parte del usuario. Esto no necesariamente convierte el uso del programa en algo sencillo o simple de realizar, puesto que de la misma manera que con R, se deben involucrar las estrategias y opciones adecuadas dentro del software para llegar al resultado obtenido.

Como se puede presuponer, las hipótesis son las mismas que se usaron en la aplicación con R y el preprocesamiento de datos se realizará partiendo del hecho de que los archivos .csv siguen siendo la fuente principal de datos.

El manejo de Orange se estructura por Widgets separados en cuatro secciones: "Data", "Visualize", "Model", "Evaluate" y "Unsupervised".

La sección "Data" contiene los widgets para cargar datos a Orange y para realizar preprocesamiento. Realizando un símil con R Studio, se podría decir que se tienen funcionalidades relativamente parecidas a las que se tenían con la biblioteca "dplyr", por lo que, además de cargar los datos, podremos escoger qué columnas son las que nos interesan para realizar las predicciones en pasos posteriores, así como generar correlaciones, unir datos, e inclusive cargar scripts de Pyhton (que no serán usados en esta aplicación). La interfaz de

Orange, así como algunas de las principales opciones de esta sección se muestran en la Figura 45:

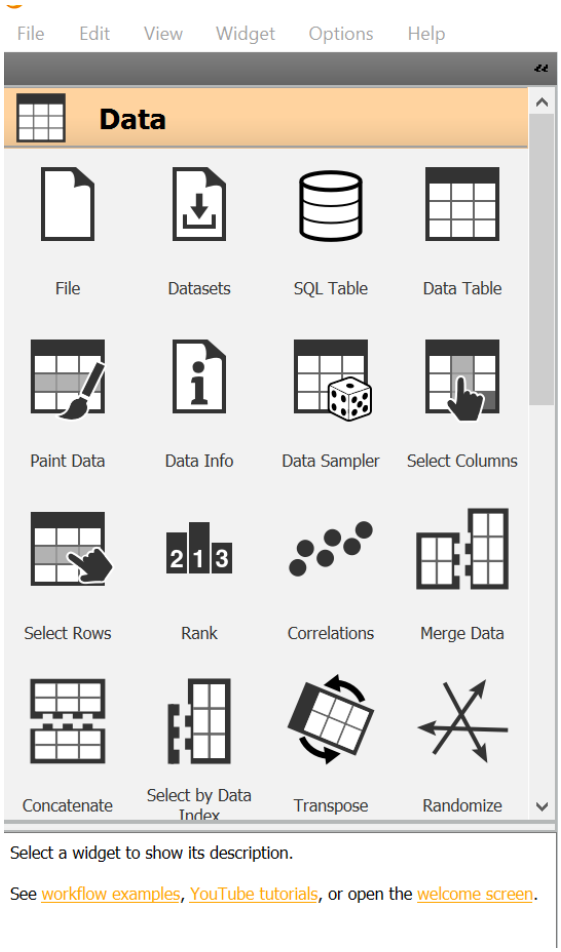


*Figura 45.I*nterfaz de Orange y opciones de la sección “Data". (Elaboración Propia)

Como se puede apreciar en la Figura 45, existen muchas maneras de importar datos, como puede ser un dataset con un tipo de archivo propio de Orange, una tabla SQL de una base de datos relacional o varios tipos de archivos como pueden ser .xlsx o .csv que es el tipo de archivo que se usará. Es importante señalar que en la sección inferior a donde se presentan los widgets, está un pequeño espacio donde se muestra una breve descripción del widget seleccionado, así como links útiles para aprender a usarlo de mejor manera.

La sección “Visualize” sirve para crear gráficos a partir de los datos cargados. Esto se puede comparar a lo visto en R Studio con la librería “ggplot”. Existen muchos tipos de gráficos que pueden ser desplegados en Orange. Esto se muestra en la Figura 46.

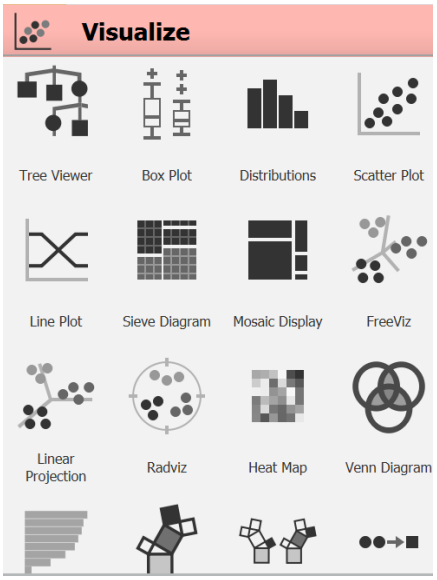


Figura 46. Interfaz de Orange. Sección "Visualize" y sus diferentes widgets. (Elaboración Propia)

La siguiente sección es "Model" que nos permite escoger un método para realizar el modelo de Minería de Datos que deseemos. Existen las opciones más comunes como Regresión Lineal, Árboles, etcétera. También se pueden guardar y cargar modelos para trabajar con más facilidad. Esto se muestra en la Figura 47.

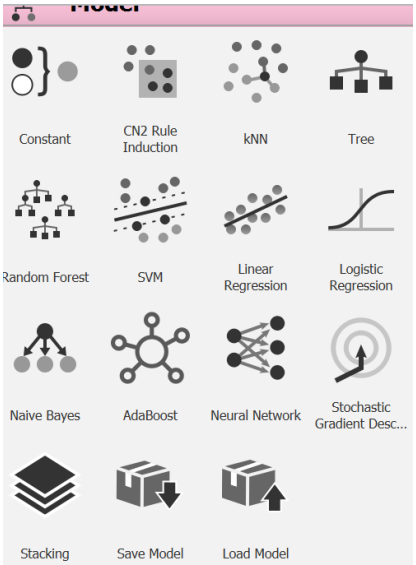


Figura 47. Interfaz de Orange. Sección "Model" y sus distintos widgets. (Elaboración Propia)

La siguiente sección "Evaluate" nos permite obtener los resultados de los modelos generados en la sección anterior, donde se puede poner a prueba el modelo, obtener predicciones, matrices de confusión, entre otras opciones. Esto se puede comparar a los resúmenes de los modelos que se obtenían en R Studio, así como las predicciones en sí mismas que se generaban; esto se muestra en la Figura 48.

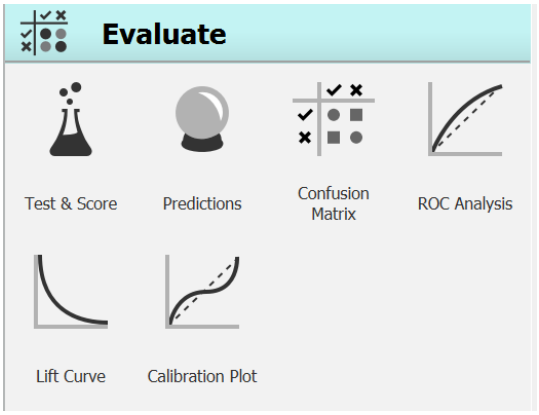


Figura 48. Interfaz de Orange. Sección "Evaluate" y sus diferentes widgets. (Elaboración Propia)

Por último, la sección "Unsupervised" nos provee algunos métodos extras basados enteramente en Aprendizaje No Supervisado, como el algoritmo K-medias. Esto se muestra en la Figura 49.

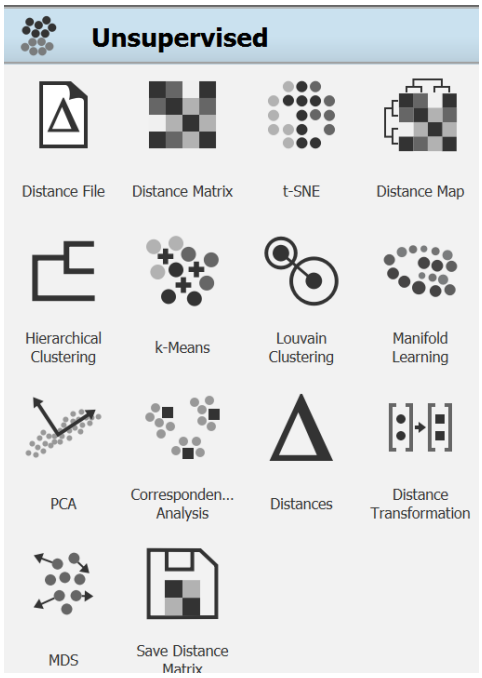


Figura 49. Interfaz de Orange. Sección "Unsupervised" y sus diferentes widgets .(Elaboración Propia)

Como se puede observar en las imágenes previas, las secciones presentes en Orange suplen en su totalidad el proceso de Minería de Datos, partiendo desde el preprocesamiento hasta la evaluación de resultados. La concatenación de estos widgets es lo que nos permitirá generar un modelo funcional con los datos que tenemos.

4.3.2 Preprocesamiento

Para comenzar el proceso se necesita cargar el archivo .csv correspondiente. Esto se hace mediante el widget "File" de la sección "Data". Simplemente se selecciona dicho widget y se lo ubica en el espacio en blanco central de la interfaz. Una vez hecho esto, al hacer doble click sobre el ícono del archivo se abre una ventana en la cual podemos elegir el archivo desde nuestro ordenador. Se busca el archivo y se lo carga.

Una de las cosas que se debe mencionar es que, a diferencia del proceso con R Studio, en este caso se debe escoger la variable objetivo desde el inicio, es decir, desde el momento mismo en que se cargan los datos. Una vez que estén subidos en Orange, se puede modificar el rol de cada variable, que por defecto estarán como "feature" y se las debe cambiar a "target". En la Figura 50 se puede observar la ventana con los datos cargados y una variable como "target" (que no corresponde a la variable objetivo que se usará en el primer modelo, solo se muestra como ejemplo). Adicionalmente, se indica el tipo de variable de cada columna, así como sus posibles valores si son texto.

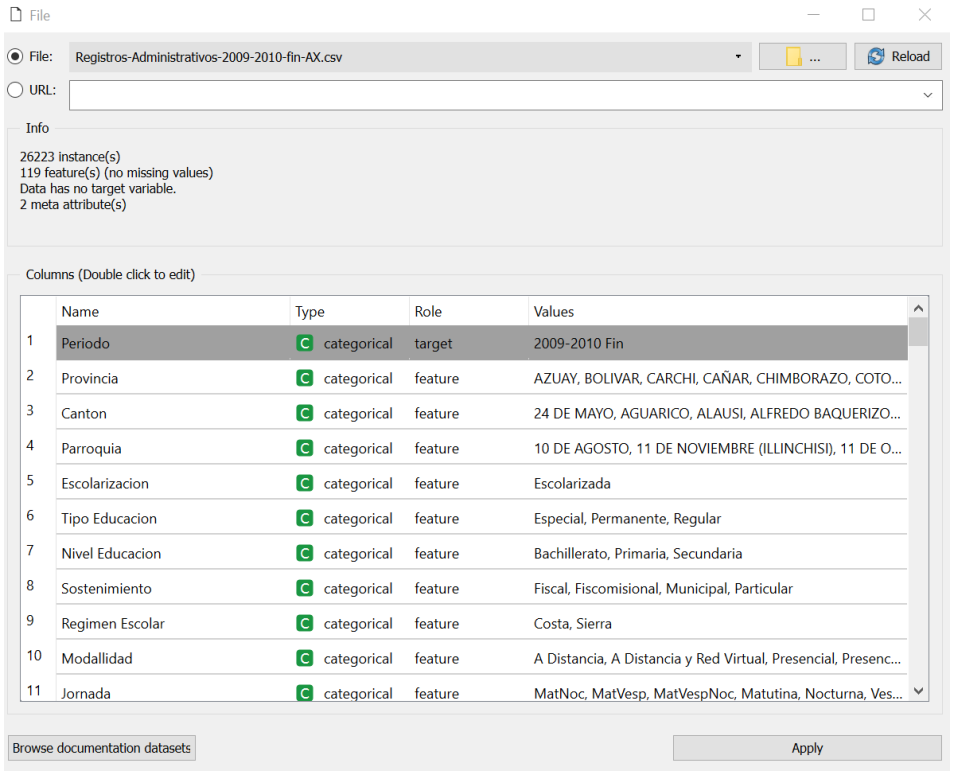


Figura 50. Ventana de Orange con la carga de datos; la variable Período se encuentra seleccionada como variable objetivo.

Una vez se haya seleccionado la variable objetivo se presiona el botón “Apply” y los datos se encontrarán correctamente cargados en Orange. Una vez que se tengan los datos cargados, lo que sigue es seleccionar las columnas que se usarán para el análisis, esto se lo hace mediante el widget “Select Columns” y para que coincida con el análisis realizado en R Studio, se procede a descartar las siguientes columnas: Período, Cantón, Parroquia, Escolarización y Modalidad (además de Nombre y Dirección de la Institución pero éstas variables por defecto quedan desafectadas al ser de tipo texto). En la ventana del widget, se procede a usar el botón de la flecha hacia la izquierda para pasar las variables que no serán usadas en el modelo, tal y como se muestra en la Figura 51, en donde también podemos ver la variable objetivo y las variables tipo texto separadas.

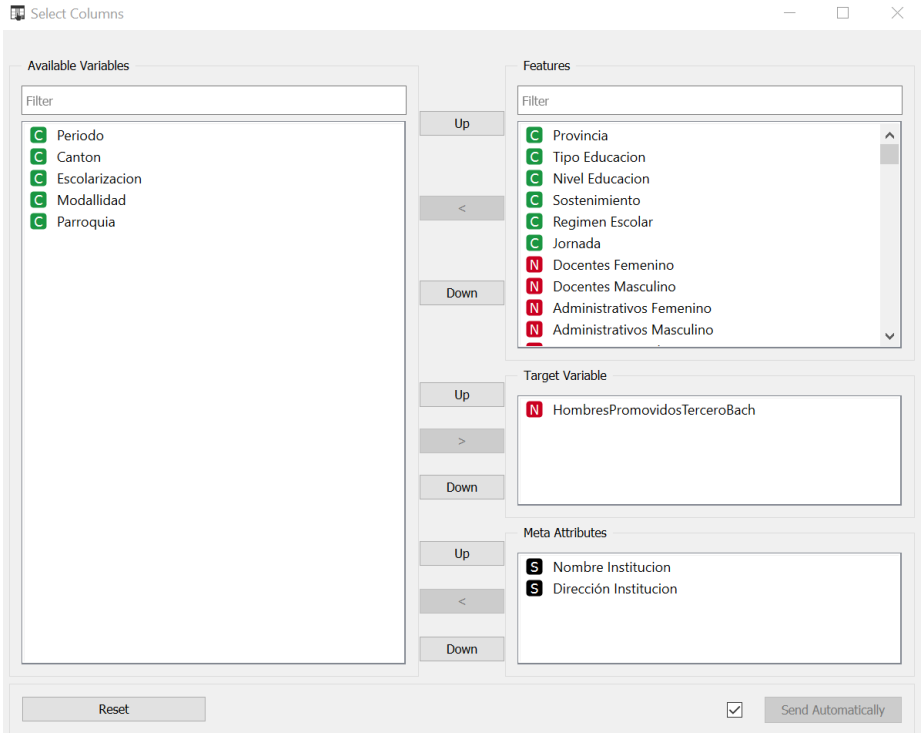


Figura 51. Selección de columnas para el modelo (a la izquierda se muestran las columnas desafectadas) en Orange. (Elaboración Propia*)*

Para llegar a la ventana que se muestra en la Figura 51, es necesario unir los dos widgets. Esta es la manera en que funciona Orange. Los widgets se comunican unos a otros y para ello se debe unir cada uno de ellos por medio de líneas de conexión, tal y como lo muestra la Figura 52. De esta forma el widget de selección de columnas sabe con qué conjunto de datos se está tratando.

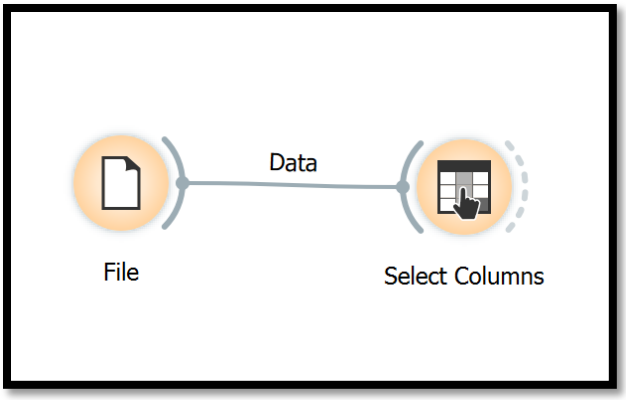


*Figura 52.*Uniendo dos widgets en Orange. (Elaboración Propia)

Los widgets no se pueden concatenar con todo el resto, muchos requieren una o más entradas específicas. Orange facilita esto especificando cuáles son las entradas y salidas de cada widget.

4.3.3 Visualización de Datos

Para la visualización de datos se hará uso de los widgets de la sección "Visualize". Como existen varios tipos de gráficos, se escogen tres para comprar su funcionalidad. Los diagramas elegidos son: Diagrama de Caja, Diagrama de Dispersión y las Distribuciones correspondientes de cada variable. No es necesario realizar esto paso a paso, puesto que se pueden concatenar los tres widgets al mismo tiempo (partiendo del widget de la selección de columnas), tal y como se muestra en la Figura 53.

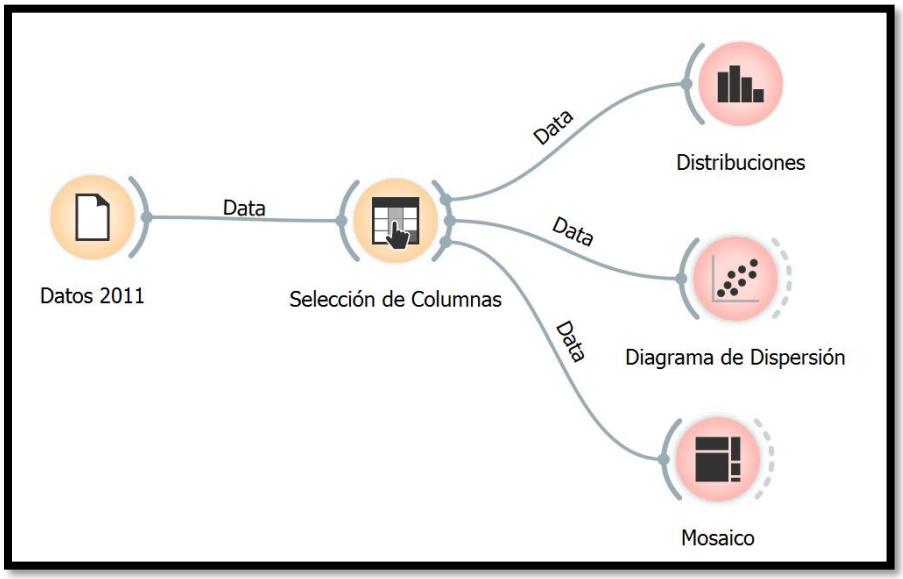


Figura 53. Concatenación de widgets de diagramas elegidos en Orange para obtener visualización de los datos. (Elaboración Propia)

Un punto importante por señalar es que se puede cambiar el nombre del widget con el fin de que sea más sencillo para el usuario conocer qué es lo que realiza el mismo (aunque el símbolo de cada widget es bastante preciso en su significado). Como se mencionó, los diagramas pueden generarse al mismo tiempo con la concatenación directa del widget que contiene los datos luego del preprocesamiento. Del mismo modo que se realizó con R, se procede a mostrar algunos gráficos que pueden mostrar cómo se encuentran los datos, para tener una idea previa antes de generar un modelo de minería.

Primero se verán las distribuciones de cada una de las variables. El widget en Orange es simple de usar y simplemente se basa en la selección de una variable y posteriormente se debe especificar otra variable para hacer el agrupamiento. Por ejemplo, en la Figura 54 se muestra la distribución de la variable Docentes Femenino agrupado por Tipo Educación. Orange separa los valores de acuerdo a colores y cuenta con una leyenda en la parte superior derecha del gráfico.

Debido a la cantidad de valores en determinadas columnas (por ejemplo, en la columna Provincia), es posible que los textos de los ejes o las leyendas se distorsionen; esto es un problema directo de la interfaz de Orange y, por lo tanto, las imágenes capturadas muestran casos en donde esto no ocurre. También se debe realizar una selección lógica del par de variables para que el gráfico resultante sea de utilidad.

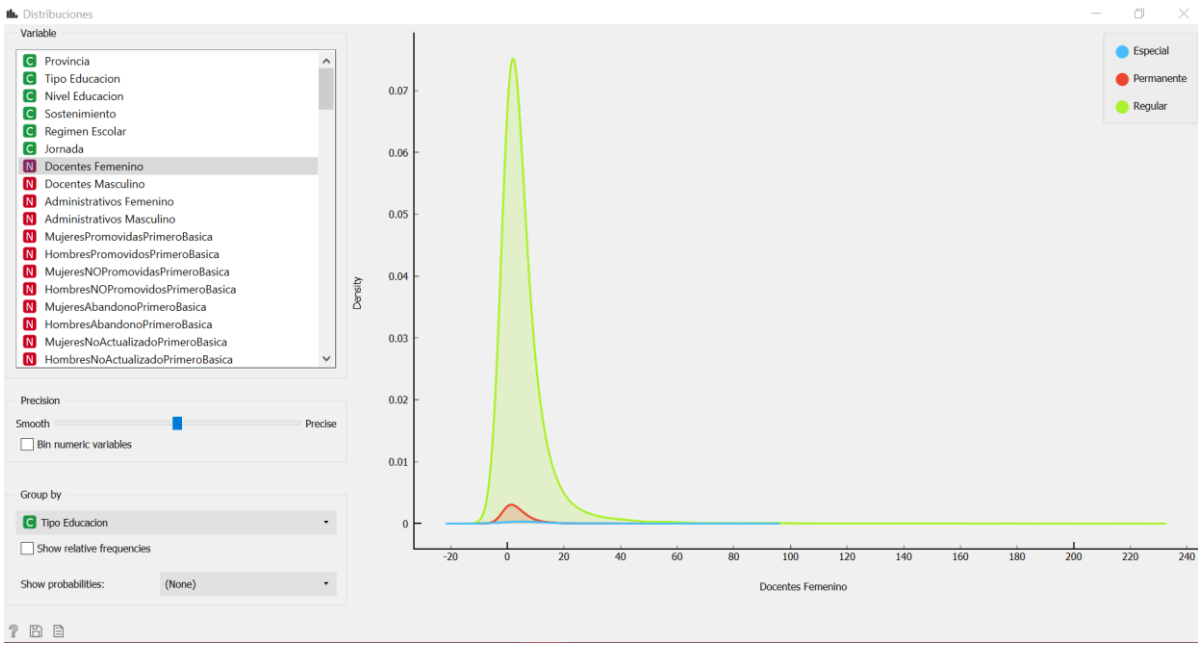


Figura 54. Distribución de la variable Docentes Femenino agrupada por la variable Tipo Educación usando Orange. (Elaboración Propia)

El siguiente tipo de diagrama es el Diagrama de Dispersión, del cual se tuvo un ejemplo previo en la aplicación en R Studio. En este caso, simplemente con el widget del mismo nombre se puede realizar algo similar. El Diagrama de Dispersión sirve para mostrar el comportamiento de dos o más variables en un grupo de datos y con ello podemos observar su relación. En la Figura 55 se puede apreciar un ejemplo con un gráfico con los ejes Hombres Promovidos Segundo de Básica y Docentes Femenino y en el gráfico se genera una especie de subdivisión por medio de figuras para representar la Costa y la Sierra; y por medio de colores para representar el tipo de sostenimiento del Plantel.

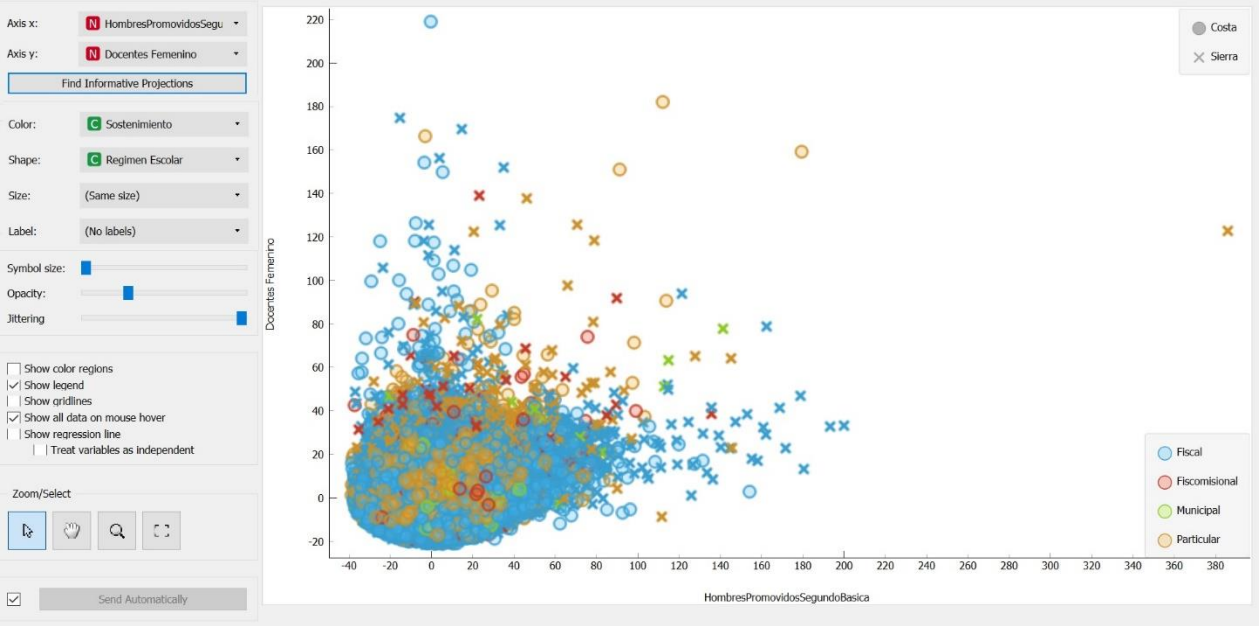


Figura 55. Ejemplo de Diagrama de Dispersión en Orange. (Elaboración Propia)

El último tipo de diagrama es el denominado “Mosaico”. Éste tipo de diagrama muestra información de ciertas variables en un mismo gráfico y separadas por figuras geométricas, usualmente rectángulos. Al igual que el Diagrama de Dispersión, con el Mosaico se puede encontrar relaciones entre variables de forma gráfica, pero de manera un poco más ordenada, debido a que solamente trabaja con rectángulos. Para construirlo se eligen dos variables principales de las cuales se desee conocer el número de datos de una variable en otra. Por ejemplo, en la Figura 56 se eligen las variables Docentes Femenino y Régimen Escolar. Se generan subgrupos en forma de rectángulos dependiendo del número de docentes mujeres en cada uno de los tipos de régimen (en este caso Costa o Sierra). Los subgrupos generados son: menos de dos, entre dos y cuatro, entre cuatro y ocho y más de ocho. Adicionalmente, se puede elegir una tercera variable para complementar el diagrama, en el caso de la Figura 56, es la

variable de Jornada. Por lo que, los cuadrados ahora se pintan de un determinado color de acuerdo a la Jornada. Entonces, se puede ver que, tanto en Costa como en Sierra, la mayoría de docentes mujeres se encuentran en planteles con Jornada Matutina (representada con color naranja).

Del mismo modo, se pueden crear distintos gráficos para comparar la situación de ciertas variables. Estos gráficos son más concisos que los vistos previamente, aunque no detallan por completo el número de especímenes en cada variable.

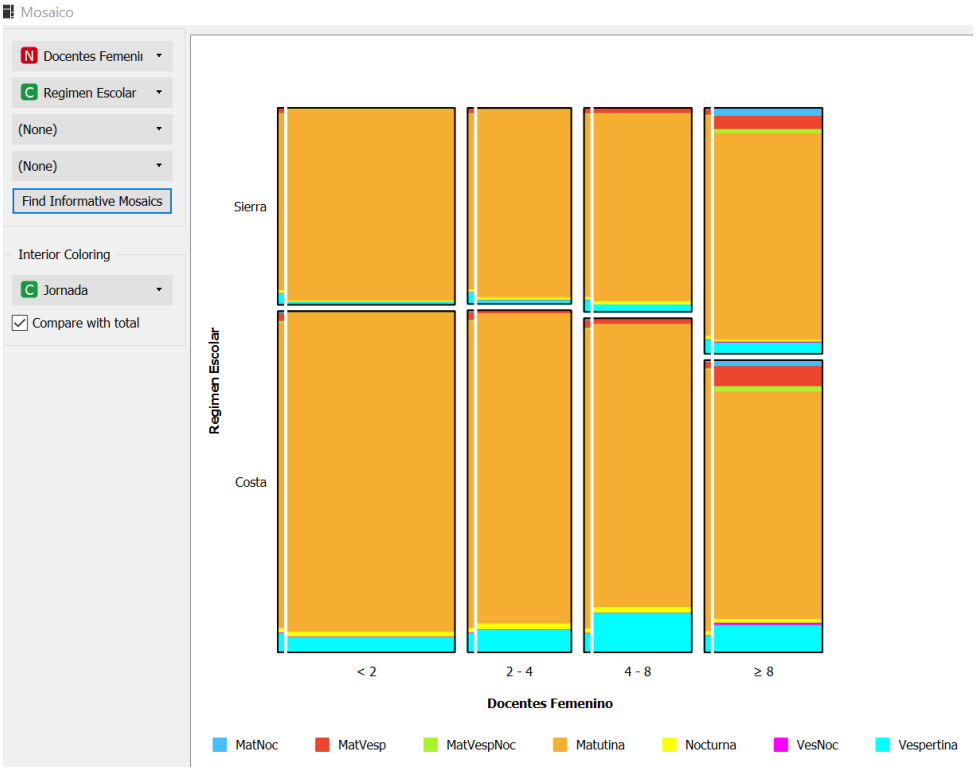


Figura 56. Diagrama Mosaico en Orange. (Elaboración Propia)

Los gráficos elegidos y los mostrados en el presente trabajo, son solo una muestra de todos los diagramas que Orange ofrece, sin embargo, con el fin de visualizar los datos, los tres tipos de diagramas elegidos son suficientes.

4.3.4 Minería de Datos (Modelado y Evaluación de Datos)

Debido a que se seguirá usando el método de Regresión Lineal y éste es uno de los ejemplos de Aprendizaje Supervisado, se debe tener dos grupos de datos: uno de entrenamiento y uno para probar el modelo. Para visualizar mejor el proceso se dividirá visualmente los datos, sin embargo, esto es algo que Orange no puede realizar directamente. Por lo tanto, se aprovechará

que los proyectos en R Studio se mantienen para guardar archivos .csv de prueba y entrenamiento. Con esto se cumplirá el primer requisito para la aplicación de Minería.

Básicamente, se guardarán todos los archivos "test" y "train" de cada uno de los proyectos en R Studio. Esto se lo realiza mediante la función "write" tal y como se muestra en la Figura 57. Los archivos se guardarán en la misma carpeta del proyecto.

```
write.csv(test, file="test2011.csv")
write.csv(train, file="train2011.csv")
write.csv(testHombresOctavo, file="testHombresOctavo2011.csv")
write.csv(trainHombresOctavo, file="trainHombresOctavo2011.csv")
write.csv(trainMujeres, file="trainMujeres2011.csv")
write.csv(testMujeres, file="trainMujeres2011.csv")
```

Figura 57. Función write en R para guardar los archivos de prueba y entrenamiento y usarlos en Orange. (Elaboración Propia)

Una vez realizado este proceso, se puede continuar. Primero, se eliminarán todos los widgets con respecto a los diagramas/gráficos, puesto que no son esenciales para el proceso. Luego, la idea general es crear dos "flujos de trabajo", es decir dos grupos de widgets concatenados: el primero trabajando sobre los datos de entrenamiento y así permitirle a la máquina realizar predicciones acertadas, las cuales se probarán sobre los datos de prueba. Los archivos .csv que se generaron en R Studio contienen una columna adicional que los numera, esta columna debe ser descartada usando la selección en Orange.

Primero, se deben cargar los archivos como ya se indicó previamente. Luego se realiza la selección de las columnas y luego se usa el widget Data Table para la visualización de los datos. Además, se da nombres a cada uno de los widgets para no perder el hilo de trabajo. Esto se muestra en la Figura 58. No se debe olvidar de establecer una variable objetivo en la selección de datos.

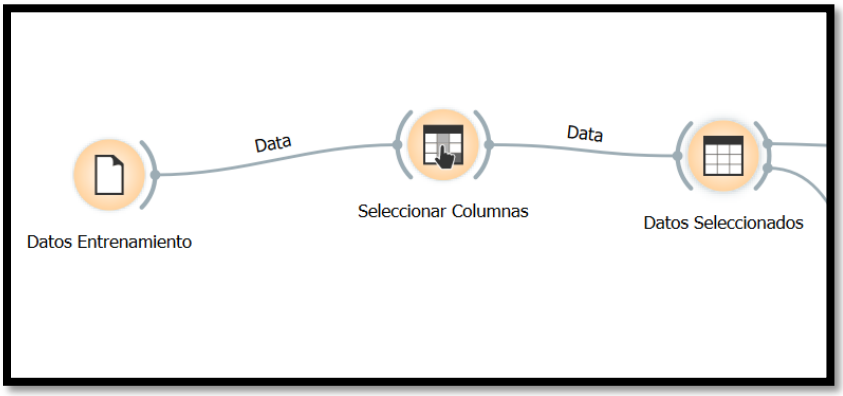


Figura 58. Carga de datos y selección de los mismos en Orange. (Elaboración Propia)

Una vez se tienen los datos seleccionados, se procede a incluir al widget de Regresión Lineal. Este widget se encuentra en la sección "Model". A partir de este widget se concatena otro widget más llamado "Predictions" que arrojará predicciones de la variable que se haya establecido como objetivo. Como se ha mencionado en previas oportunidades, la máquina trabaja sobre un conjunto de entrenamiento y arroja predicciones para un conjunto de prueba. Es en este momento en el que se juntan los dos flujos de trabajo, para obtener predicciones de los datos, tal y como se muestra en la Figura 59.

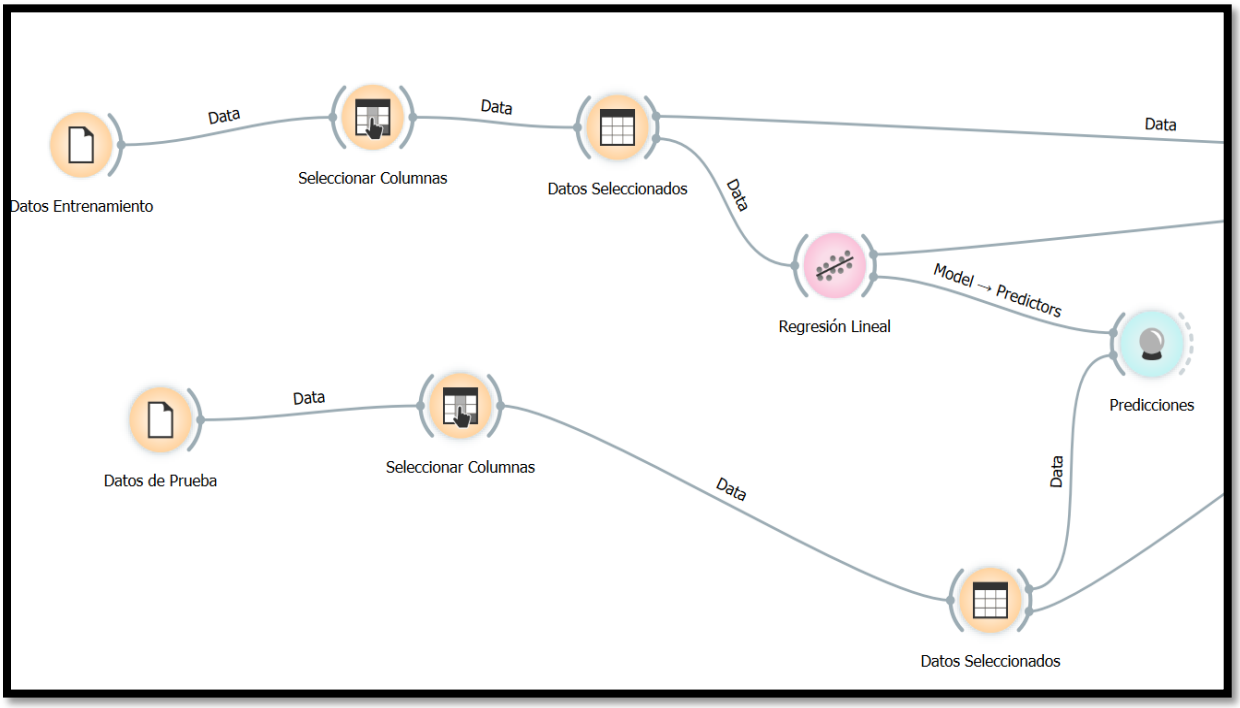


Figura 59. Generación de predicciones. Se muestran los flujos de trabajo de los datos de entrenamiento y prueba en Orange. (Elaboración Propia)

Si los datos están correctamente estructurados y la regresión lineal ha sido aplicada de manera exitosa, las predicciones se desplegarán haciendo doble click en el widget y esto se indica en la Figura 60.

Figura 60. Predicciones obtenidas con el proceso realizado en Orange. (Elaboración Propia)

Es importante señalar lo que se manda al modelo para continuar. En este caso, la regresión lineal toma como entrada a los datos seleccionados del conjunto de entrenamiento. Para las predicciones, se necesita enviar como entradas al modelo en sí y también a los datos seleccionados del conjunto de pruebas.

Debido a la gran cantidad de datos en cada uno de los subconjuntos, no se puede establecer si el modelo ha resultado exitoso o no (debido a que la comparación entre los datos reales y los predichos sería un proceso largo y engorroso). Es por esto por lo que existe un widget para comprobar los resultados del modelo llamado “Test & Score”. Este widget va a requerir como entradas los datos de prueba, entrenamiento y lo que se conoce como “aprendiz” que viene a ser el modelo como tal. El modelo completo se muestra en la Figura 61.

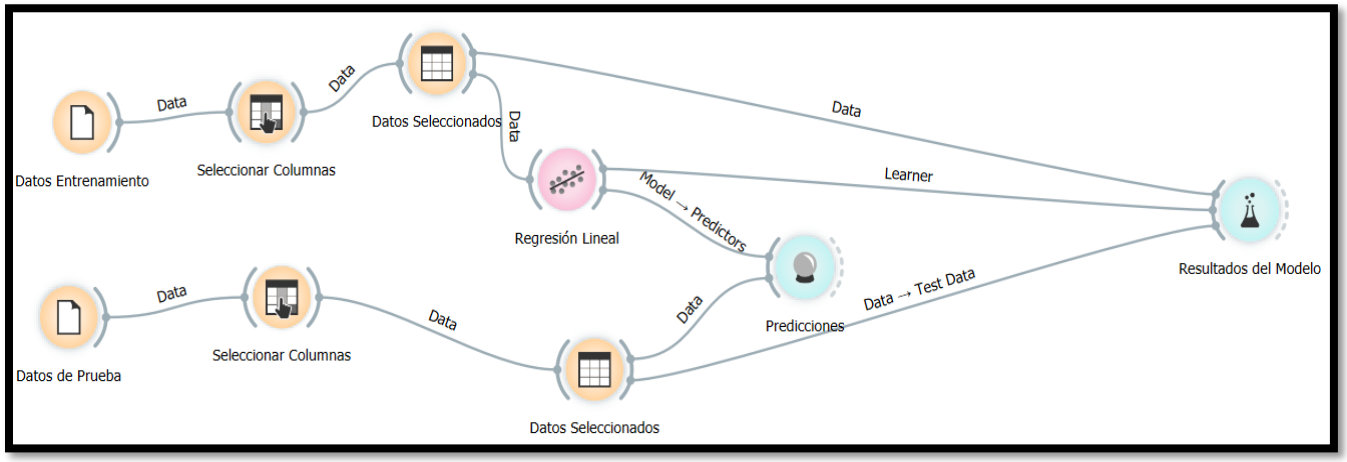


Figura 61. Modelo de Minería de Datos usando Regresión Lineal en Orange. (Elaboración Propia)

Desafortunadamente, Orange no permite la edición de las líneas de conexión entre cada widget, por lo que podría existir un problema de estética a la hora de presentar el modelo, con superposiciones en algunos casos. Como bien muestra la información de cada uno de los widgets, muchos de éstos aceptan distintas entradas, lo que indica que existen varias maneras de generar un modelo de esta clase, sin embargo, se considera una forma simple e intuitiva la que se muestra en la Figura 61, por lo que, resulta suficiente para el presente trabajo. Lo que contiene el widget de resultados es un compendio de distintas métricas con las cuales se puede evaluar el modelo, entre éstas se encuentra el ya mencionado R Cuadrado Ajustado (con el símbolo de R2). Los resultados se muestran en la Figura 62.

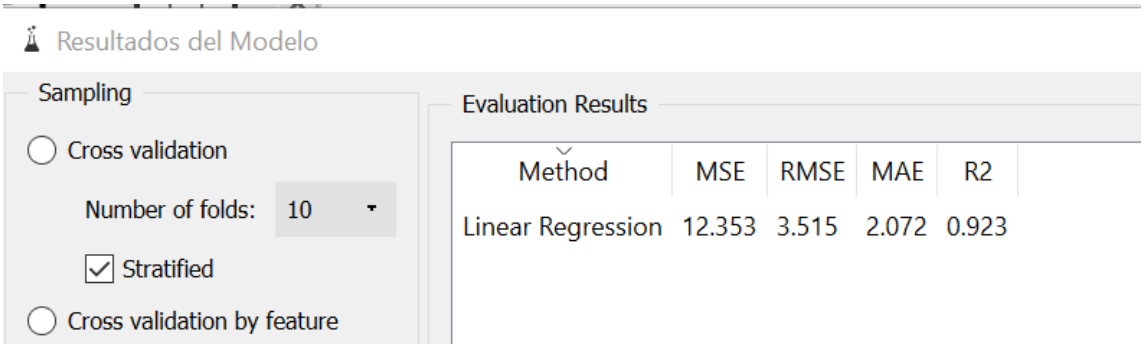


Figura 62. Resultados del Modelo de Regresión Lineal en Orange mostrado anteriormente. (Elaboración Propia)

Lo que se puede ver es que el RCA está en un 91.9% un valor similar a los obtenidos en R Studio y que se considera como aceptable.

Usando el mismo archivo, se puede poner en práctica uno de los métodos de Aprendizaje No Supervisado, las Redes Neuronales. Se podría trabajar con el mismo flujo de trabajo, sin embargo, se debe recordar que los métodos de Aprendizaje No Supervisado no requieren que la máquina sea entrenada. Por lo tanto, para las Redes Neuronales se puede usar directamente el archivo .csv total de cada período lectivo. El modelo viene a ser el mismo que el visto previamente, solo que para realizar las predicciones se usan únicamente los datos fuente y no se tiene una separación, es decir, en ningún momento la máquina está aprendiendo. Los resultados del modelo, según su RCA, son menos efectivos que los obtenidos con Regresión Lineal. Esto se puede apreciar en las Figuras 63 y 64.

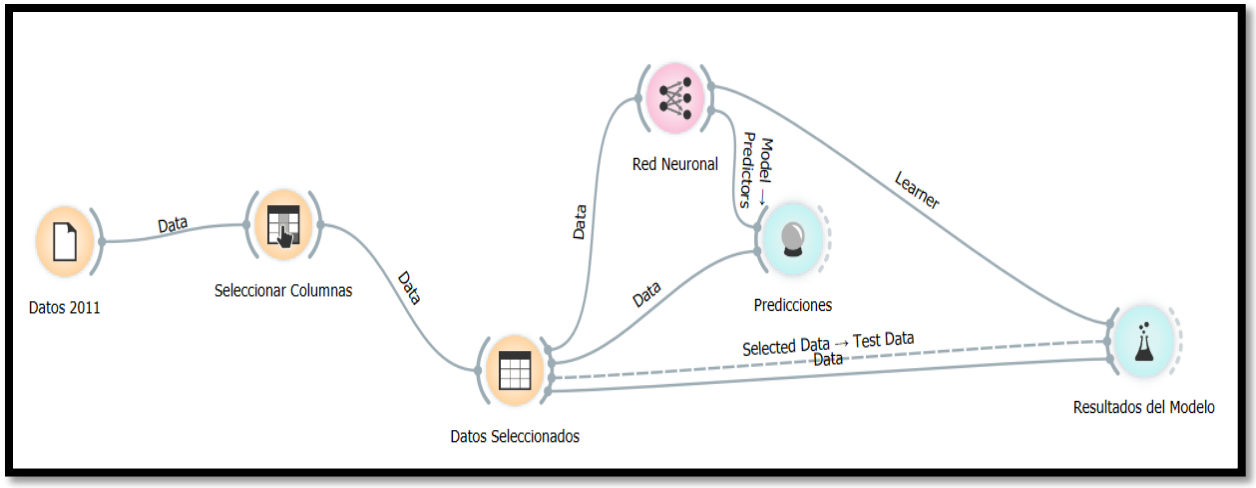


Figura 63. Modelo usando Redes Neuronales en Orange. (Elaboración Propia)

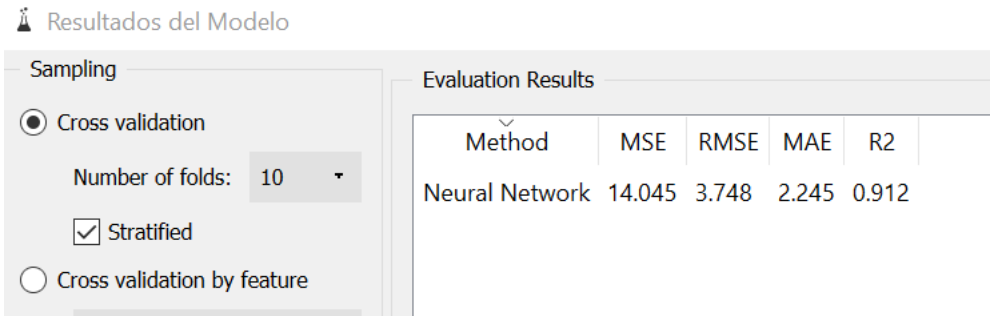


Figura 64. Resultados del Modelo con Redes Neuronales en Orange. (Elaboración Propia)

Un punto interesante a tener en cuenta es que, no es necesario usar dos conjuntos de datos para realizar el proceso, Orange puede manejar la separación de los datos mediante la "Validación Cruzada". El hecho de tener un modelo con dos flujos de trabajo es solamente para volver aún más gráfico el proceso. Un modelo de este estilo, más reducido se puede apreciar en la Figura 65, sin embargo, no debe confundirse con el modelo empleado con Redes Neuronales, donde no se debía particionar los datos.

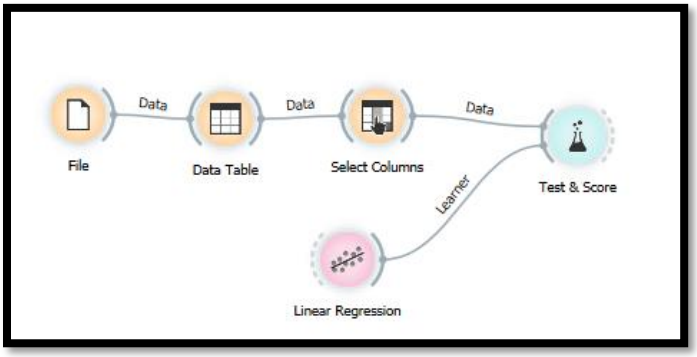


Figura 65. Modelo reducido sin la necesidad de particionar los datos en Orange. (Elaboración Propia)

La opción de Validación Cruzada se encuentra en el widget “Test & Score” y solamente debe tener con entrada una selección de datos y el método elegido para el modelo.

Una de las opciones más interesantes que se podía obtener con R Studio en el resumen del modelo era la relación o la influencia de ciertas variables para la variable objetivo establecida. En el caso de Orange, esto se puede replicar haciendo uso del widget “Rank”, que establece una clasificación de todas las variables que pueden llegar a afectar a la variable objetivo señalada, esto basándose en un algoritmo de filtro en base a las características o variables de un conjunto de datos llamado RReliefF, que viene a ser una mejora de otros algoritmos similares que toman como base a uno denominado simplemente “Relief”. Una vez conectado el widget con una selección de datos, en cualquier modelo, se procede a dejar marcada únicamente la casilla de RReliefF y en una tabla aparecen ordenadas las variables con más peso (el número al lado de cada variable solamente significa la cantidad de posibles valores que pueden tomar dentro del conjunto general de datos), tal y como se muestra en la Figura 66.

Figura 66. Ranking de las variables según el algoritmo RReliefF con la variable Hombres Promovidos Segundo Básica como variable objetivo en Orange. (Elaboración Propia)

Una vez que se ha logrado obtener un modelo equiparable a lo realizado en R Studio, se procede a realizar las mismas pruebas que en la aplicación anterior, es decir, eliminar variables del modelo para ver si éstas afectan o no al RCA cuando se analicen los resultados. Por supuesto, se deben generar los modelos correspondientes para las otras dos hipótesis que han quedado pendientes. En general, no existe ningún inconveniente en el hecho de usar distintos flujos de trabajo con los mismos archivos base (en el mismo proyecto de Orange), sin embargo, el área de trabajo quedaría muy grande, por lo que se generarán proyectos para cada hipótesis, teniendo 3 archivos por cada período lectivo.

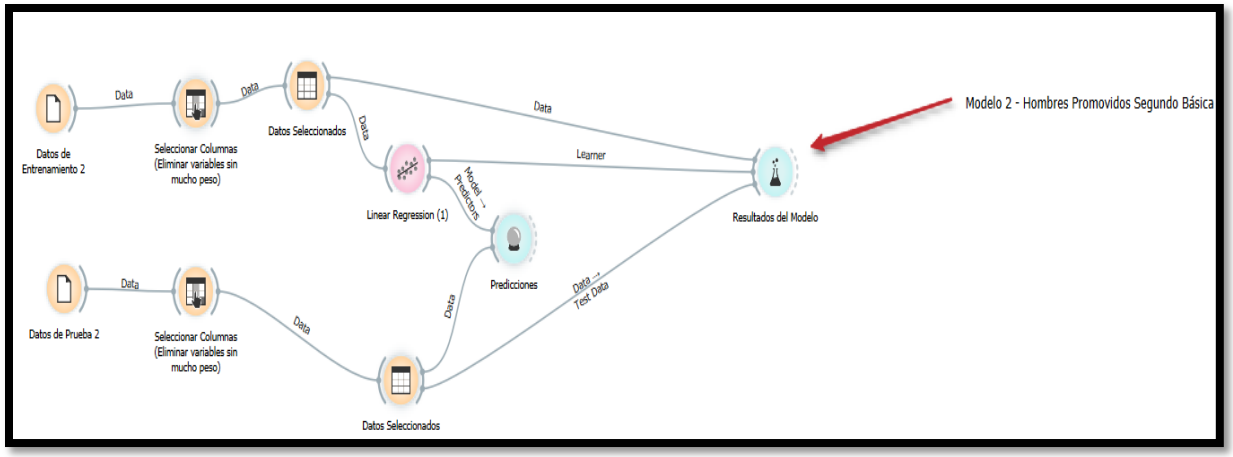


Figura 67. Segunda Versión del Modelo con la variable Hombres Promovidos Segundo Básica como objetivo en Orange. (Elaboración Propia)

En la Figura 67 se aprecia una segunda versión del modelo con la variable objetivo respectiva, eliminando algunas columnas de la selección para comprobar si éstas influyen en el resultado del RCA obtenido. También se aprecia una flecha y un texto aparte del modelo que sirven para mantener un orden interno, aclarando qué representa cada flujo de trabajo. El resultado es similar al obtenido en un inicio con un RCA de 92.2%, con lo que se infiere que las variables eliminadas realmente no tenían impacto en el modelo.

Para el caso de las mujeres que abandonan sus estudios del tercer año de Bachillerato, se construye un modelo similar que se aprecia en la Figura 68. Los resultados de este modelo son similares a los obtenidos en la mayoría de scripts en R, con un RCA de 73.3%. Además, usando el widget Rank se obtiene que las dos variables más influyentes para este modelo son la Jornada (es decir si la institución imparte clases de mañana, tarde o por la noche, o la combinación de éstas) y Sostenimiento (es decir, si la institución es particular, fiscal, fiscomisional o municipal).

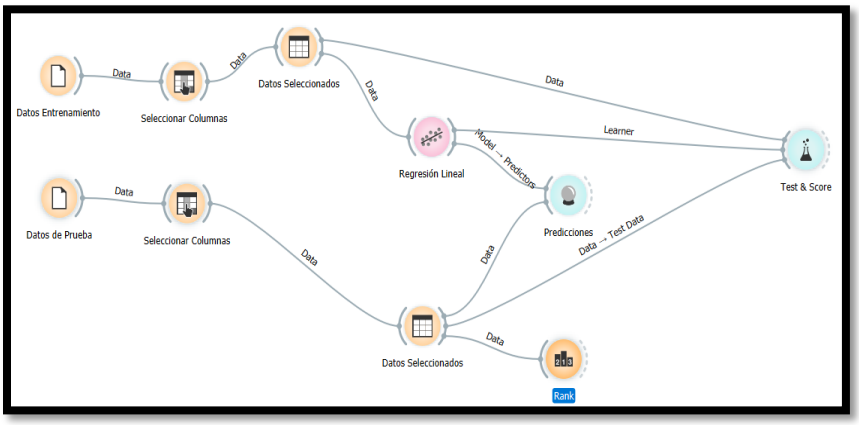


Figura 68. Primer modelo realizado con la variable Mujeres Abandono Tercero Bach como objetivo en Orange. (Elaboración Propia)

A continuación, se eliminan las variables menos influyentes del modelo según el ranking realizado para ver si existe algún cambio en el modelo. Se insiste en que, la eliminación de variables no implica una mejora en las métricas de evaluación del modelo, sin embargo, su inclusión puede estar inflando artificialmente los valores de las métricas, confundiendo al usuario sobre la eficacia de su modelo. Es por esto por lo que, se procede a la eliminación de estas variables y si, el RCA se mantiene o decae ligeramente (hasta 1 punto porcentual), quiere decir que estas variables no tenían ninguna clase de impacto y que, por lo tanto, no necesitan incluirse en el modelo final.

Para el nuevo modelo se van a eliminar todas las variables correspondientes a los estudiantes de Primaria (es decir, desde Primero de Básica hasta Séptimo de Básica). Esto provoca una disminución considerable del RCA cayendo hasta el 48.5%, indicando que algunas de las variables eliminadas, sí resultaron importantes para la construcción del modelo. Aunque, lo interesante de este modelo es que el ranking nos indica que la variable Provincia es la más influyente.

Como se obtuvo un mal RCA, se pueden probar otros métodos como las Redes Neuronales antes aplicadas, o también se puede usar otro método como el conocido Random Forest o Bosques Aleatorios (otro método de Aprendizaje Supervisado derivado de los Árboles de Decisión). Usando este método, el RCA que se obtiene es de 91.6%, y se lo realizó manteniendo las variables iniciales, lo que indica una mejora en el modelo.

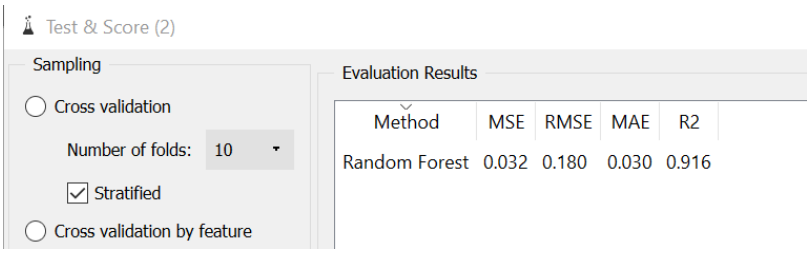


Figura 69. Resultados modelo de Mujeres Abandono Tercero Bach como variable objetivo usando Random Forest en Orange. (Elaboración Propia)

Tomando como base al último modelo realizado con Random Forest, el ranking de las variables se muestra en la Figura 70 y nos presenta tres aspectos importantes, pero que terminan por desestimar la hipótesis planteada.

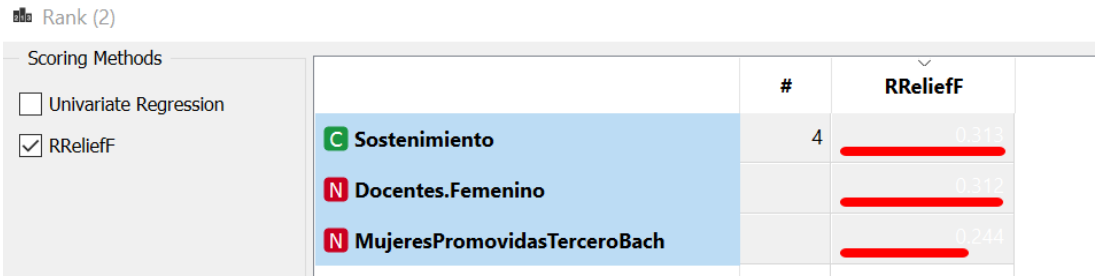


Figura 70. Ranking de las variables según el algoritmo RReliefF con la variable Mujeres Abandono Tercero Bach como objetivo y usando el método Random Forest en el modelo en Orange. (Elaboración Propia)

Aclarando que, en el primer modelo se obtuvo una gran influencia de la variable Provincia, pero dicho modelo puede considerarse como más impreciso que el último mostrado.

Para trabajar con los datos de la última hipótesis se realizará una variante que involucra el tipo de modelo simplificado visto previamente. Esto se aplica con el fin de observar cómo Orange puede presentar varios resultados de distintos métodos al mismo tiempo. La idea general es trabajar con la opción de Validación Cruzada para evitar particionar los datos y también para trabajar con Redes Neuronales que no necesitan de este aprendizaje. Este modelo se aprecia en la Figura 71.

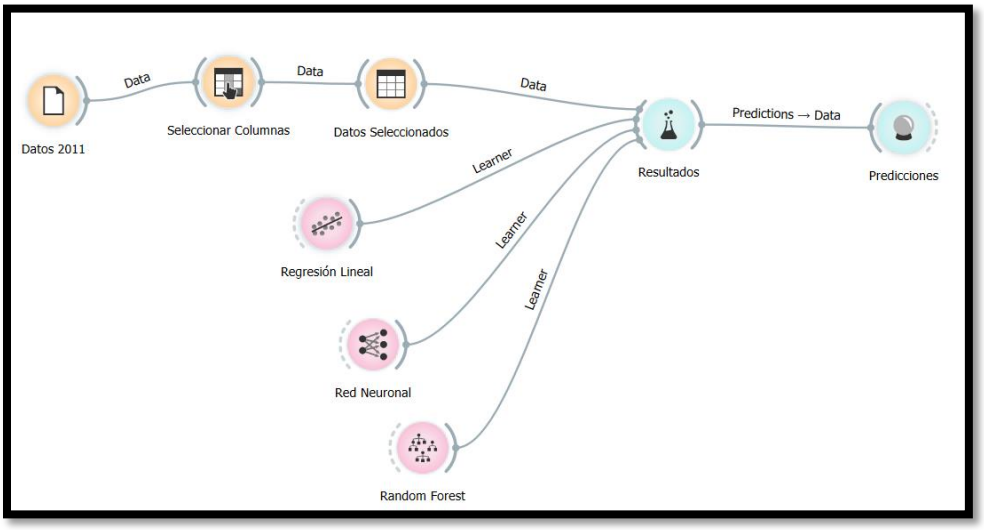


Figura 71. Modelo simplificado usando varios métodos en Orange. (Elaboración Propia)

En este caso se emplean los métodos de Regresión Lineal, Redes Neuronales y Random Forest. El flujo de trabajo consiste en subir la data y elegir las columnas para el análisis y posteriormente con el widget denominado en la Figura como "Resultados" preparar los datos para aplicar los modelos. Las predicciones se obtienen a partir del widget Resultados, no directamente de los modelos como en ocasiones pasadas.

Los resultados demuestran que aplicar Redes Neuronales resulta más efectivo con un RCA de 73.8%. Esto se aprecia en la Figura 72 y se entiende que los datos en este caso no terminan de ajustarse para alguno de los métodos elegidos. Además, usando el widget de Rank se encuentra que la variable Sostenimiento sí llega a ser influyente, sin embargo, también se incluyen variables que no resultan lógicas en este caso, como Mujeres Abandono Primero de Básica. Lo que se procede a realizar es eliminar algunas de estas variables para comprobar si realmente afectan al modelo.

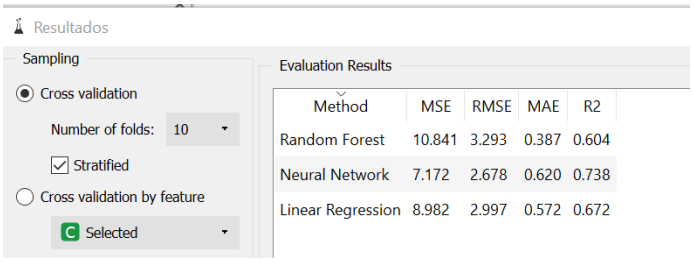


Figura 72. Resultados del modelo realizado en Orange usando tres tipos distintos de métodos con la variable Hombres NO Aprobados Octavo Básica como objetivo. (Elaboración Propia)

Eliminando algunas variables, los resultados se mantienen para Random Forest y Redes Neuronales, aunque con Regresión Lineal se tiene un aumento hasta alcanzar un RCA de 73.6%, mientras que en el ranking se termina confirmando que, en efecto, la variable de Sostenimiento resulta importante en conjunto con la variable de Provincia, es decir, se podría concluir que existe una relación real entre el número de estudiantes que reprueba el octavo año de educación básica y el tipo de sostenimiento del plantel educativo, así como la provincia a la que pertenecen dichas instituciones. Estos datos se aprecian en la Figura 73.

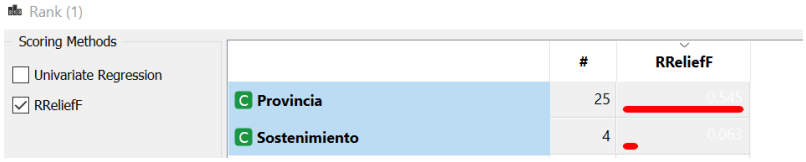


Figura 73. Ranking de las variables según el algoritmo RReliefF con la variable Hombres NO Aprobados Octavo Básica como objetivo en Orange. (Elaboración Propia)

Por facilidad, el resto de proyectos se realizarán con la versión simplificada del modelo, en donde no solamente el flujo de trabajo se disminuye y no se requiere la carga de dos archivos de datos por cada flujo, sino que también se puede incluir distintos métodos para tener una rango más amplio de opciones sobre las cuales elegir el mejor modelo posible. Las repercusiones de los datos obtenidos en este capítulo serán tratadas en el siguiente, realizando comparaciones sobre lo arrojado por ambas herramientas en cuanto a las predicciones y a los RCA obtenidos en cada modelo y en cada año lectivo.

5. Análisis Comparativo de Herramientas

5.1 Evaluación previa de resultados obtenidos con las herramientas seleccionadas

5.1.1 Comparación Previa del trabajo con las herramientas seleccionadas.

Una vez concluido todo el trabajo con las herramientas R Studio y Orange, es necesario evaluar lo obtenido. Con ambos programas se crearon distintos archivos para cada período lectivo, en el caso de Orange se separaron dichos archivos en uno para cada hipótesis planteada previamente, con el fin de mantener un orden interno para tener rápido acceso a los archivos.

El proceso de Minería de Datos aplicado se basó en lo planteado en el capítulo 2 con respecto a este tema. La definición y exploración de los datos se los trató en el capítulo 3 así como también parte del preprocesamiento y preparación de los datos; sin embargo, en cada proyecto desarrollado con las herramientas seleccionadas, se establecieron distintas secciones que corresponden a: Preprocesamiento, Visualización, Modelado y Evaluación. Esto se realizó debido a que la etapa de Visualización consistió en analizar los datos mediante distintos gráficos y para ello se necesitaba que la data se encuentre formateada de modo que haya sufrido todos los cambios que el Preprocesamiento demandaba.

En cuanto al Modelado, el proceso se basó en la utilización de métodos de Aprendizaje Supervisado y No Supervisado populares como Regresión Lineal, que es el método principal de todos los proyectos realizados en R Studio, pero también se exploraron algunas soluciones adicionales como el uso de Redes Neuronales para el caso de algoritmos de Aprendizaje No Supervisado y algunos otros métodos de Aprendizaje Supervisado como el denominado "Random Forest". El proceso de evaluación empezó con los resúmenes de los modelos en R Studio y el ranking de variables en Orange, así como las respectivas métricas que arrojaron los modelos realizados. Evidentemente, esta evaluación debe continuar en esta sección para poder llegar a conclusiones tangibles.

Una de las razones para utilizar no una sino dos herramientas para el proceso de Minería de Datos es que se pueden complementar los resultados para corroborar o mejorar distintos aspectos, como las métricas obtenidas (en este caso la métrica principal con la que se realizarán las comparaciones será el denominado "R Cuadrado Ajustado" muchas veces abreviado como R2, R^2 o simplemente RCA.

El trabajo desarrollado con R Studio es tradicional y se asemeja a otros software que permiten analizar datos como Anaconda Python, que sirve justamente como un medio open-source para Aprendizaje Supervisado y Ciencia de Datos. Se basa en la construcción de un script en un entorno amigable con distintas subsecciones que permiten aligerar la carga de trabajo, por ejemplo, mantener el editor de texto separado de la visualización de variables y la visualización de gráficos generados. Una de las características más importantes del trabajo con R Studio es la facilidad para agregar nuevas bibliotecas, con lo que simplemente el programador debe conocer los comandos o funciones relativas a cada biblioteca para poder trabajar. En el presente estudio se utilizaron hasta tres bibliotecas distintas para generar satisfactoriamente los modelos de Minería de Datos, todas ellas de libre acceso y sin ninguna restricción. El trabajo con R se ajusta mucho a lo que se pretendía llevar a cabo, por lo que la decisión de incluir esta herramienta está justificada.

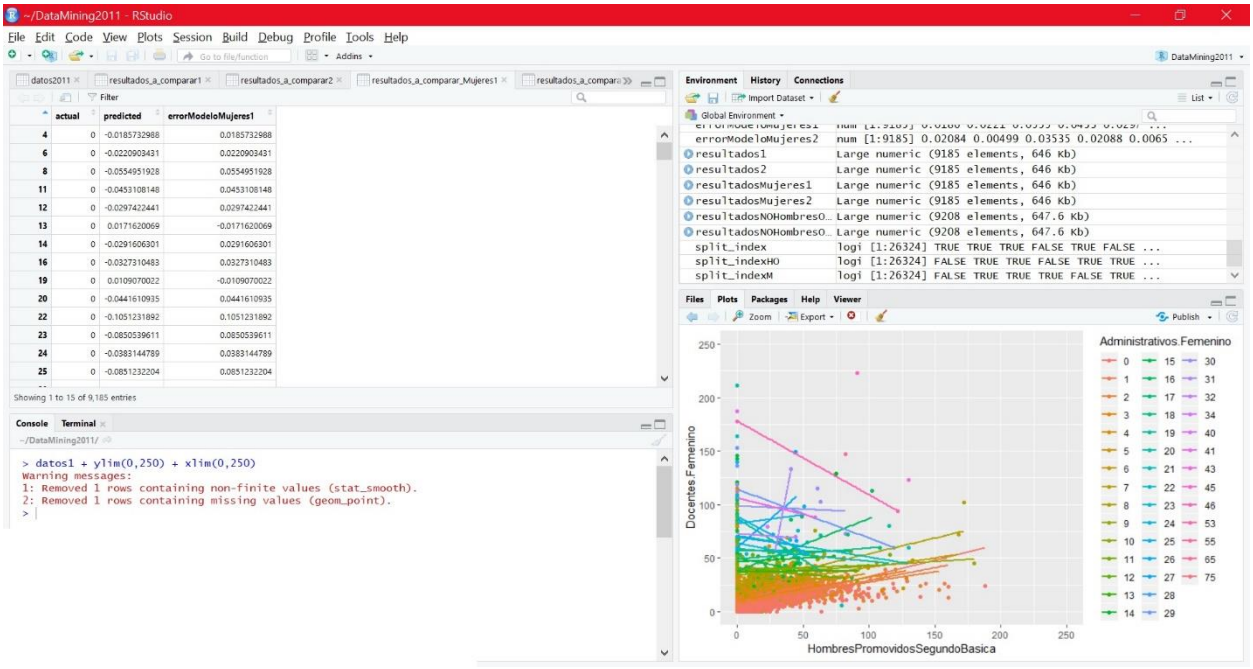


Figura 74. Trabajo con R Studio, donde se muestran sus respectivas subsecciones. (Elaboración Propia)

En cuanto a Orange, existen muchas maneras de replicar el trabajo para seguir el proceso general de Minería de Datos, sin embargo, es radicalmente distinto. Primero, Orange trabaja con programación visual, es decir, no se debe escribir líneas de código para que la máquina realice alguna acción. Lo que se hace con Orange es crear "flujos de trabajo" mediante distintos "widget" que vienen a ser pequeños círculos con una función determinada, este flujo de trabajo termina asemejándose a un grafo, pero también representa cómo funciona la programación más

tradicional puesto que se hacen llamados a distintas funciones para poder completar algún objetivo más grande. Orange está basado en Python, así que una de las opciones más llamativas para programadores es que se pueden cargar scripts en este lenguaje para agilizar el trabajo y/o añadir nuevos algoritmos.

Sin embargo, si se habla de los algoritmos o métodos para llevar a cabo un modelo, se debe mencionar que Orange cuenta con una selección limitada de éstos y que muchas de sus variantes no pueden ser aplicadas. Esto podría ser un obstáculo para trabajos más profundos con la data, por lo que siempre se prefiere algún medio más tradicional para trabajar, aunque esto no significa que se haya tenido alguna dificultad o limitante con este software, sino más bien el mismo ha servido para demostrar una manera alterna de cómo llevar a cabo un mismo trabajo.

Un aspecto interesante por destacar es que, no se necesitó realizar ningún ajuste demasiado complejo en Orange a los conjuntos de datos que previamente fueron preprocesados y manejados con R. El único punto a tener en cuenta es que se debe establecer la variable objetivo desde el inicio para poder decirle al programa que se trabajará de esa manera. El resto, es decir, la Visualización, el Modelado y parte de la Evaluación se realizan con los distintos widgets que están disponibles, concatenando unos con otros hasta crear un flujo de trabajo lógico que permita obtener resultados satisfactorios. La elección de Orange también se puede considerar un acierto.

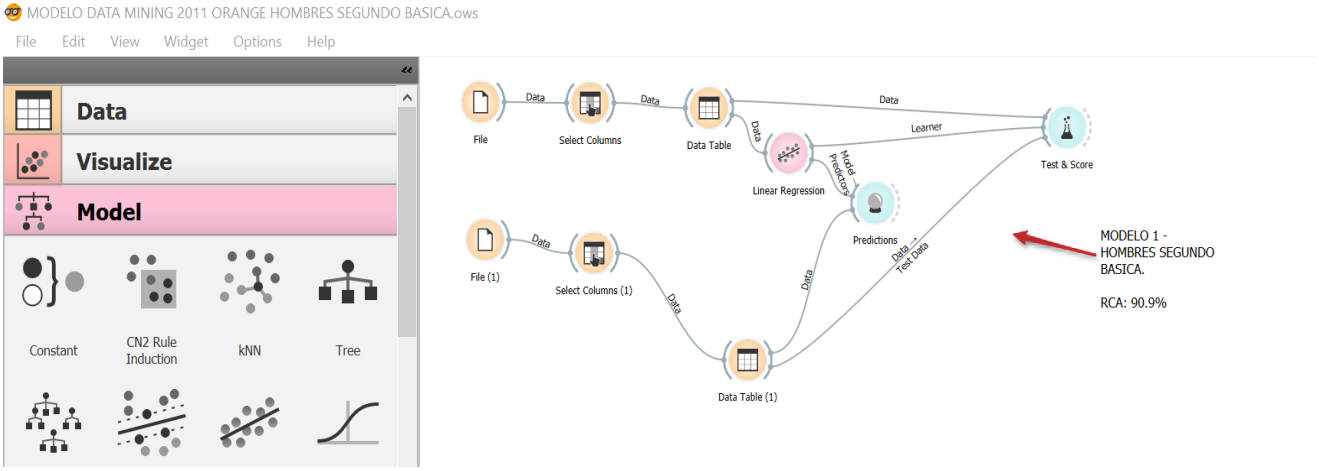


Figura 75. Trabajo con Orange, donde se muestra la concatenación de widgets para construir un modelo. (Elaboración Propia)

Vale aclarar que los datos, como se mencionó en los capítulos 3 y 4, son archivos .csv que fueron transformados desde archivos .xlsx que se encontraban alojados en repositorios del Ministerio de Educación y son de libre acceso a la ciudadanía.

5.1.2 Resultados previos obtenidos de cada hipótesis.

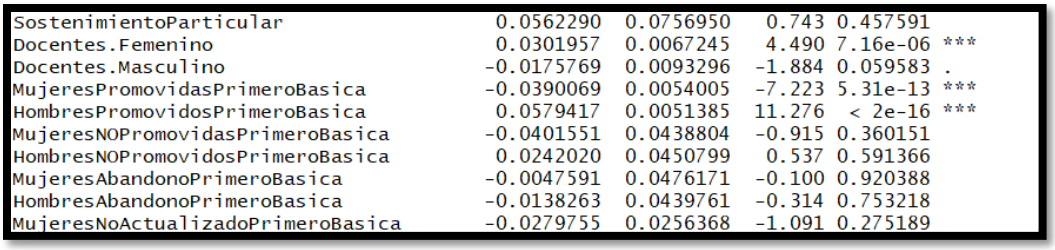
Este análisis se separará en tres puntos principales, uno para cada una de las hipótesis planteadas en el capítulo 4, esto debido a que se realizaron distintos modelos para cada caso y se pueden comparar los resultados obtenidos.

5.1.2.1 Hombres Promovidos Segundo de Básica

La hipótesis relacionada con esta variable hacía alusión a una posible relación entre el número de estudiantes promovidos del Segundo Año de Educación Básica y el número de docentes mujeres de cada institución, así como el tipo de jornada en la que se encuentran estudiando. Como se entiende, el proceso de Minería de Datos principal arrojó predicciones mediante un modelo con un método en concreto, la manera en que se pueden contestar las hipótesis es mediante la valoración de cada software para determinar qué variables influyeron en el modelo. Aunque es importante responder las hipótesis, no será el principal aspecto a tener en cuenta para valorar un modelo, sino serán las métricas obtenidas, que se evaluarán posteriormente.

Los modelos con esta variable son constantes en todos los períodos lectivos a la hora de señalar que la variable Docentes.Femenino es relevante para la construcción del modelo, con lo que se podría señalar que existe una relación directamente proporcional entre estas dos variables. Por otro lado, dentro del proceso que se siguió, para comprobar si las métricas del modelo eran correctas, se procedió a eliminar algunas variables que no se mostraban como importantes en las primeras versiones de los modelos con esta variable. Lo que sucedió fue que las métricas no decayeron, con lo que se entiende que no eran importantes; entre esas variables eliminadas se encontraba la variable de Jornada. Esto cuando se habla del trabajo desarrollado en R Studio.

Con respecto a Orange, no se eliminó la variable de Jornada, sino todas las variables de los estudiantes de Bachillerato y Preparatoria que realmente no tenían relación. Lo que se obtiene es que en la mayoría de modelos es que las variables más influyentes son las de los estudiantes varones promovidos de los siguientes niveles de Primaria, aunque se mantiene una importancia media para la variable de Docentes Femenino. Esto puede hacer alusión a que existe una proporción similar entre los estudiantes que aprueban cada uno de los niveles de Primaria.



SostenimientoParticular	0.0562290	0.0756950	0.743	0.457591	
Docentes.Femenino	0.0301957	0.0067245	4.490	7.16e-06	***
Docentes.Masculino	-0.0175769	0.0093296	-1.884	0.059583	.
MujeresPromovidasPrimeroBasica	-0.0390069	0.0054005	-7.223	5.31e-13	***
HombresPromovidosPrimeroBasica	0.0579417	0.0051385	11.276	< 2e-16	***
MujeresNOPromovidasPrimeroBasica	-0.0401551	0.0438804	-0.915	0.360151	
HombresNOPromovidosPrimeroBasica	0.0242020	0.0450799	0.537	0.591366	
MujeresAbandonoPrimeroBasica	-0.0047591	0.0476171	-0.100	0.920388	
HombresAbandonoPrimeroBasica	-0.0138263	0.0439761	-0.314	0.753218	
MujeresNoActualizadoPrimeroBasica	-0.0279755	0.0256368	-1.091	0.275189	

Figura 76. Resultados de uno de los modelos con la variable Hombres Promovidos Segundo Básica como objetivo en RStudio, donde se aprecia la relevancia de la variable Docentes.Femenino en el modelo. (Elaboración Propia)

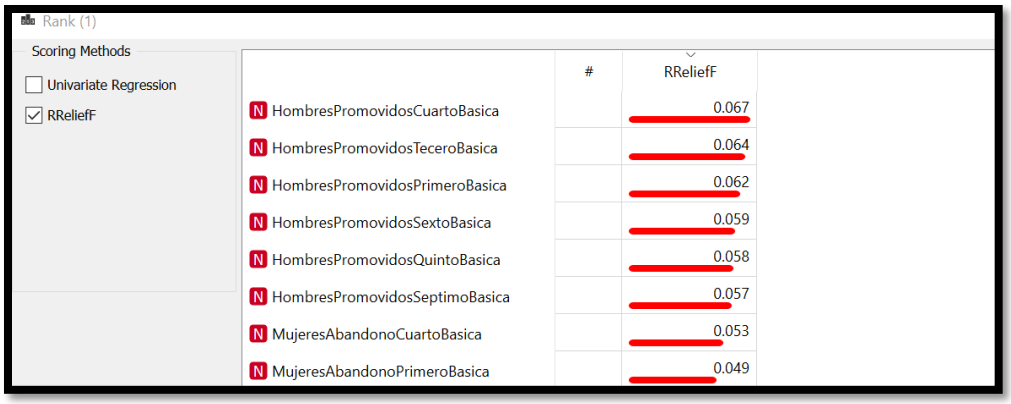


Figura 77. Resultados de uno de los modelos con la variable Hombres Promovidos Segundo Básica como objetivo en ORANGE, donde se aprecia la relevancia de las variables de hombres promovidos de los distintos años de educación en Primaria. (Elaboración Propia)

A pesar de que en este caso existe una relativa coincidencia entre los valores obtenidos tanto en R como en Orange, el hecho es que pueden variar debido a que no usan el mismo algoritmo para realizar el "ranking" de las variables, sin embargo, esto no es motivo para que existan resultados muy distintos, solamente debería variar la influencia (grande o media) de las principales variables, que ya debieron ser identificadas.

5.1.2.2 Mujeres Abandono Tercero de Bachillerato

La hipótesis relacionada con esta variable hacía alusión a una posible relación entre el número de estudiantes mujeres que abandonaban sus estudios mientras cursaban el Tercero de Bachillerato y si las instituciones donde estas estudiantes cursaban sus estudios se encontraban en la Costa o en la Sierra. Esta hipótesis se plantea por el conocimiento popular (un conocimiento de "boca a boca" y sin ningún estudio que lo sustente) de que existe un mayor número de estudiantes mujeres que abandonan sus estudios debido a que existe un mayor número de embarazos juveniles. En este caso, cuando se repasa el análisis de cada modelo respectivo a esta variable usando R Studio, se obtiene que efectivamente la variable Provincia es influyente en el modelo, sin embargo, en este caso, no es suficiente dado que, se necesita una especificación

en cuanto a qué provincias son las que más representantes tienen. Por lo tanto, se realiza un conteo del número de estudiantes mujeres que abandonaron sus estudios de cada provincia. El análisis se realiza tomando en cuenta que, la Costa del Ecuador se considera conformada por las siguientes provincias: Esmeraldas, Manabí, Guayas, Santa Elena, Los Ríos, El Oro y Santo Domingo de los Tsáchilas; mientras que la Sierra del Ecuador se conforma por Carchi, Imbabura, Pichincha, Cotopaxi, Tungurahua, Chimborazo, Bolívar, Cañar, Azuay y Loja.

Lo que se obtiene puede dar pie a establecer una posible comprobación de la hipótesis, es decir, que existen más mujeres en la Costa del país que abandonan sus estudios en el Tercero de Bachillerato que las estudiantes en la Sierra. Sin embargo, esto solo es un indicio, pues se debería realizar un análisis de las hipótesis para comprobarlas; dicho análisis no está comprendido en este trabajo. Los datos fueron extraídos directamente de los archivos .csv y los gráficos mostrados en la Figura 78 se generaron directamente en Excel para no depender de una de las dos herramientas seleccionadas. Se puede notar una relación de 3 a 2 aproximadamente en todos los años de los períodos lectivos analizados, las gráficas corresponden a los años 2011 y 2017 (períodos 2010-2011 y 2016-2017).

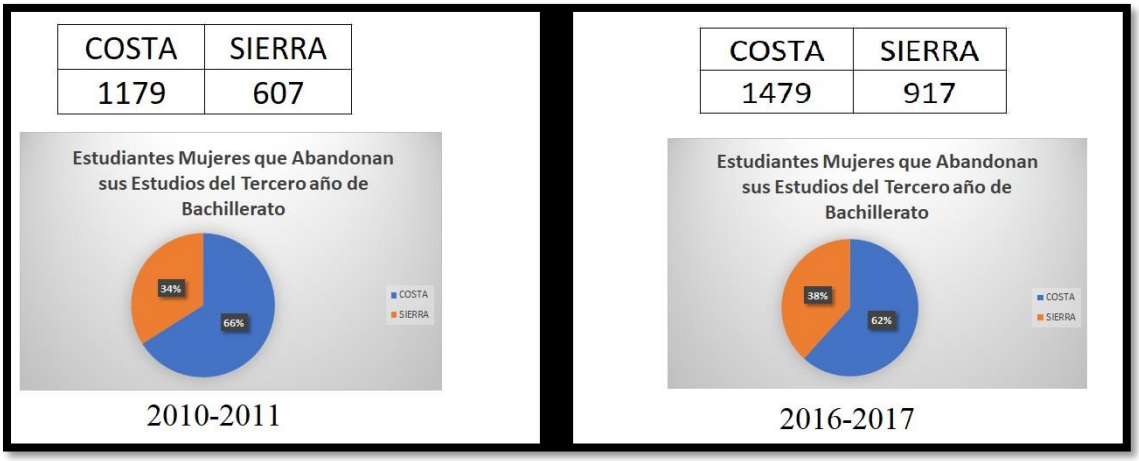


Figura 78. Comparación entre dos períodos lectivos del número de estudiantes mujeres que abandonan sus estudios al cursar el Tercero de Bachillerato. (Elaboración Propia)

Cuando se realiza el proceso respectivo usando Orange, se llega a la conclusión de que sí existe una relación con la variable Provincia, pero como ya se mencionó, esto resulta insuficiente. Un aspecto por destacar es que la variable "Régimen Escolar" no puede servir para el análisis en el caso de que hubiese sido determinante, debido a que el régimen no necesariamente quiere decir que una institución está o no en la Costa del país. Por ejemplo,

existen instituciones en Azuay, una provincia de la Sierra, que llevan su régimen académico con la modalidad Costa.

Es comprensible que exista alguna persona que esté en contra de lo que se ha realizado aquí, debido a que no es necesario llevar a cabo un análisis de Minería de Datos para llegar a la conclusión obtenida. Si bien es cierto esto, la cantidad de datos a considerar es muy grande y el tiempo que implicó realizar este análisis sin ninguna herramienta enfocada en Minería, solamente las opciones básicas de Excel, es mucho mayor que el tiempo invertido en la generación de un modelo en Orange o R para obtener estas respuestas. El hecho de que se haya usado este análisis posterior con los gráficos tipo pastel fue única y exclusivamente para esclarecer un punto que las herramientas ya habían indicado, pero que era necesario especificar.

Por último, es necesario mencionar que, este conjunto de datos, el de las mujeres que abandonan sus estudios en Tercero de Bachillerato, fue el que más problemas arrojó con los modelos y sus métricas (como se verá próximamente), debido a que existen muchos planteles educativos que cuentan con cero estudiantes que abandonaron. Al tener la mayoría de datos en cero en toda la columna (salvo algún dato disperso que termina por ser definitivo), las herramientas obtienen predicciones muy cercanas a cero, pero que cuentan con algunos decimales de más que a la postre terminan afectando en sus métricas. Evidentemente, el trabajo consistió en buscar la manera de elevar el valor de estas métricas hasta obtener modelos eficientes y en base a estos, acceder a las variables influyentes.

5.1.2.3 Hombres No Promovidos Octavo de Básica

Como premisa, se establece una posible relación entre el número de estudiantes varones que reprueban el octavo año de Educación Básica y el sostenimiento del plantel educativo (es decir, si el plantel es fiscal o particular). La hipótesis surge de otro conocimiento popular, sin basarse en estudios reales, que dice que la educación privada es mucho más benevolente con los estudiantes que en una institución pública, por lo que hay más estudiantes reprobados en colegios/escuelas públicas. No se hace una distinción entre fiscal, fiscomisional o municipal, pues a la final son instituciones no privadas, sin embargo, se dejó en todos los archivos a las instituciones de todas las clase de sostenimientos para mantener el orden original.

Lo obtenido aquí es que existe efectivamente una influencia de la variable Sostenimiento, en R Studio se puede apreciar cómo existe esta influencia en los modelos generados. Aunque existen algunas excepciones, con algunos períodos lectivos donde esta variable aparece como no

tan influyente, en la mayoría de casos se tiene que la variable de Sostenimiento sí afecta. En los modelos desarrollados con esta variable como objetivo, no se eliminó nunca la variable Sostenimiento y se obtuvieron resultados aceptables.

En la Imagen 5 se aprecia la influencia en uno de los resúmenes del modelo desarrollado en R Studio, la imagen corresponde a los datos del período lectivo 2013-2014 y muestra cómo la variable Sostenimiento tiene gran influencia en su opción "Fiscomisional".

```
SostenimientoFiscomisional      -1.945e-01  4.985e-02  -3.901 9.62e-05 ***
SostenimientoMunicipal           1.966e-02  8.797e-02   0.224 0.823141
SostenimientoParticular         -4.862e-02  2.660e-02  -1.828 0.067623 .
```

Figura 79. Resultados de uno de los modelos con la variable Hombres NO Promovidos Octavo Básica como objetivo en R Studio, donde se aprecia la relevancia de la variable Sostenimiento en el modelo. (Elaboración Propia)

Es importante señalar que R Studio sí provee una descripción más exacta de qué opción o valor de las variables son las que resultan representativas en el modelo. Como se aprecia en la Figura 79, únicamente el valor "Fiscomisional" tiene tres estrellas, mientras que el valor "Particular" solo cuenta con un punto que representa una baja influencia en el modelo; esto es algo que se pierde en Orange, pero en este caso puntual no resulta un problema.

Para el caso de Orange se tiene un poco más de constancia con respecto a los resultados, teniendo en casi todos los modelos de los períodos lectivos con esta variable como objetivo, una muestra de que la variable Sostenimiento sí es la más influyente (o de las más influyentes) para la construcción del modelo. Esto puede deberse, como se dijo previamente, a la ligera diferencia que se presenta entre ambas herramientas al tener dos algoritmos base para realizar el ranking de variables. En Orange se obtienen resultados en donde la variable Sostenimiento comparte lugar con la variable Provincia o Jornada, que son otras muy representativas en todos los modelos realizados.

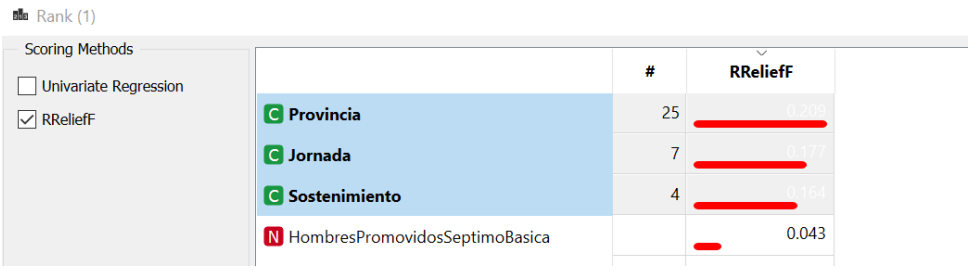


Figura 80. Resultados de uno de los modelos con la variable Hombres NO Promovidos Octavo Básica como objetivo en ORANGE, donde se aprecia la relevancia de la variable Sostenimiento. (Elaboración Propia)

La variable Sostenimiento no siempre tiene el primer lugar en el ranking que ofrece Orange, sin embargo, siempre está en el "podio". En la Figura 80 se aprecia cómo, se encuentra

en el tercer lugar de importancia, luego de Provincia y Jornada, según el algoritmo especificado. Debido a que esto se repite para la mayoría de modelos, se encuentra una tendencia que ayudaría en un posible análisis de la hipótesis para comprobarla.

5.2 Establecimiento de parámetros para evaluar resultados.

Una vez que se especificaron los resultados para los modelos con respecto a las variables influyentes, es necesario tratar el tema de las métricas, que es la mejor forma de entender cómo se comporta un modelo y si es necesario descartarlo o no.

La primera parte de este análisis tuvo como parámetro principal la constancia de los valores a través de los años en los períodos lectivos correspondientes. A pesar de que esto tiene que ser valorado, para evaluar las predicciones y los modelos en sí mismos, se deben anexar nuevos criterios.

Durante el proceso de Minería de Datos realizado con R Studio, uno de los puntos que se trató fue el de obtener los errores de cada una de las predicciones realizadas, así como también el Error Cuadrático Medio. Esto debe ser replicado para los datos obtenidos con Orange para poder realizar las comparaciones en el siguiente punto del presente trabajo. Por último, la métrica con la que se trabajará será el R Cuadrado Ajustado, que a grandes rasgos lo que establece es un valor para medir qué tan bien se ajusta el modelo para los datos y adopta valores entre cero y uno, mientras más cerca esté de uno, más preciso será el modelo.

Con los errores de cada predicción, el Error Cuadrático Medio y el R Cuadrado Ajustado se conforman las bases para la comparación respectiva de este análisis. Además, se insiste en la importancia de observar la constancia de los resultados a través de los períodos lectivos para generar las conclusiones correspondientes.

5.3 Determinación de tasas de error de cada resultado mediante validación cruzada.

Los errores de cada predicción fueron obtenidos en cada uno de los modelos realizados con R, al contrario que con el trabajo con Orange, en donde únicamente se cuenta con los datos netos arrojados por las predicciones de los modelos. Para comenzar, se exportarán los errores obtenidos con R por medio de la función “write.csv” que ya fue usada previamente, dicha función crea un archivo csv y dado que en todos los modelos realizados ya se contaba con un conjunto de datos con los valores reales, las predicciones y los errores, el trabajo que queda es

exportar dicho conjunto. En la Figura 81 se muestra un resumen dela tabla con estos datos, correspondiente al modelo con la variable Hombres Promovidos Segundo Básica como objetivo y perteneciente al período lectivo 2016-2017.

	Valores Reales	Predicciones	Error
1	0	1.552460754	-1.552460754
5	1	0.897013728	0.102986272
6	13	10.46542105	2.534578951
7	19	17.56320574	1.436794257
8	3	3.913072064	-0.913072064
10	0	2.114264269	-2.114264269
12	0	1.8264626	-1.8264626
20	0	-1.408987911	1.408987911
27	0	2.228040108	-2.228040108
29	3	2.906394287	0.093605713
30	0	1.290783888	-1.290783888
31	1	0.240983323	0.759016677
32	0	-0.054578037	0.054578037
34	0	0.989958676	-0.989958676
37	2	1.450723152	0.549276848
42	0	0.591930433	-0.591930433
45	1	1.645415532	-0.645415532

Figura 81. Tabla con las predicciones y errores obtenidas con uno de los modelos del período lectivo 2016-2017 usando RStudio. (Elaboración Propia)

La numeración que se muestra en la primera fila a la izquierda corresponde a la numeración real de los datos del archivo original de ese período lectivo, en donde ciertas filas de datos fueron descartadas por no cumplir con las condiciones impuestas al realizar el preprocesamiento de los datos, por lo que la numeración va a tener saltos. Este proceso se realiza con todos los proyectos en R correspondientes. Adicionalmente, se puede copiar el valor del error cuadrático medio en el mismo archivo creado para su posterior análisis. El error cuadrático medio (MSE por sus siglas en inglés) mide el promedio de los cuadrados de los errores, en este caso se obtuvo de manera manual aplicando una fórmula. Es otra métrica usada para el análisis de modelos basados en regresión.

Una vez que se ha completado la exportación de datos provenientes de los modelos en R, se debe realizar el mismo proceso con los archivos de Orange. Para el trabajo con Orange se necesita extender el flujo de trabajo con unos widgets adicionales que permitan crear un archivo .csv con los resultados de las predicciones, tal y como se muestra en la Figura 82.

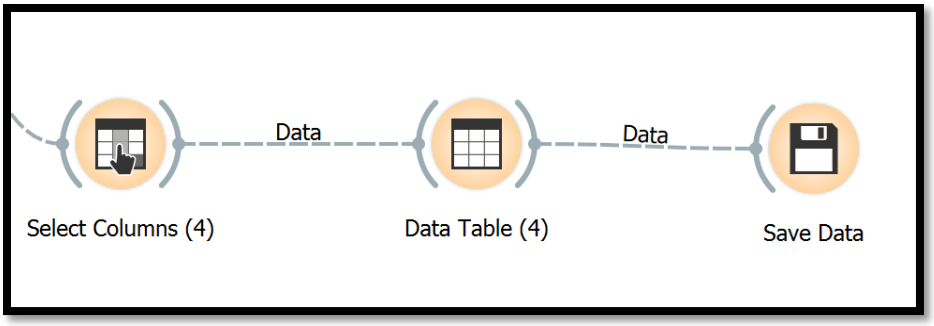


Figura 82. Extensión del flujo de trabajo en cada modelo realizado en Orange para poder exportar las predicciones resultantes. (Elaboración Propia)

El widget para seleccionar columnas debe concatenarse con el widget Predicciones en el caso en que se trabaje separando los datos en conjuntos de entrenamiento y prueba, mientras que se debe concatenar con el widget Resultados si se trabaja directamente con el repositorio original de cada período lectivo. En el widget de selección de columnas se deben ubicar todas las variables en la parte izquierda y en la parte superior derecha se deja la variable objetivo acompañada de las predicciones arrojadas por cada método usado en el modelo. Esta selección de columnas no afecta en nada al modelo, solamente se usa para que al crear el archivo .csv no se generen más datos de los necesarios. Por lo tanto, en el apartado "Available Variables" se ubican todos los datos que no se requieren en el archivo a crear, las secciones de "Target Variable" y "Meta Attributes" se quedan vacías y en la sección "Features" se ubican los datos necesarios para los archivos. Esto se puede observar en la Figura 83.

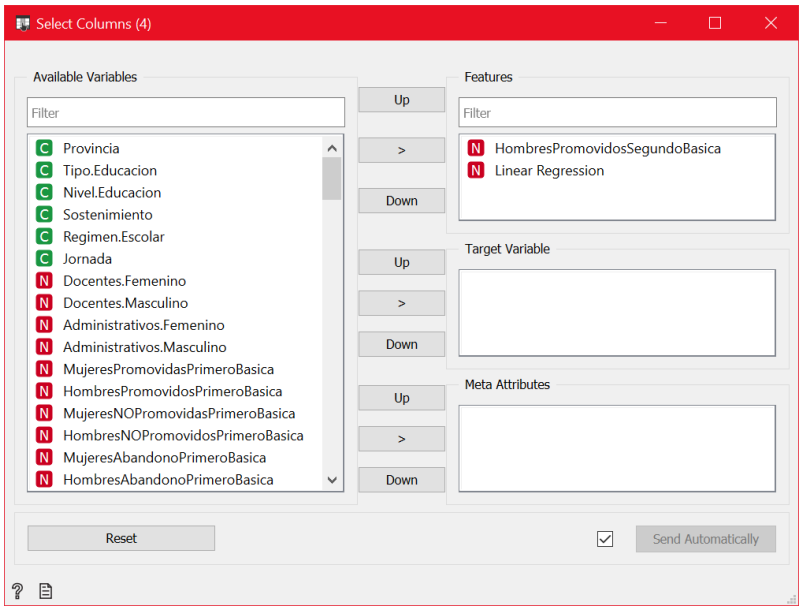


Figura 83. Selección de variables en Orange para su exportación en un archivo de extensión csv. (Elaboración Propia)

El caso de la Imagen 9 corresponde al proyecto relacionado con la variable Hombres Promovidos Segundo Básica del período lectivo 2016-2017.

Una vez estén seleccionados los datos, se concatenan con el widget "Data Table" para que estos datos seleccionados se conformen como un nuevo conjunto, el cual se concatena respectivamente con el widget "Save Data". Para guardar el archivo se debe dar doble click sobre el widget, aunque para futuras ocasiones si se deja habilitado el guardado automático, esto no será necesario puesto que cada vez que se concatene este widget se abrirá una ventana emergente para seleccionar el destino para guardar el archivo.

Si se siguieron los pasos descritos se tendrá un archivo con únicamente la columna de los datos originales de la variable respectiva y un número determinado de columnas adicionales que representan las predicciones de cada método usado. Posteriormente, se procede a generar el error absoluto, restando el valor original del valor predicho y obteniendo su valor absoluto. Se crean nuevas columnas (Error RL = Error Regresión Lineal; Error RN= Error Red Neuronal; Error RF = Error Random Forest). Asimismo, se obtiene el error absoluto medio de cada método. En cuanto a los errores cuadráticos medios, así como los R Cuadrado Ajustado se procede a observar en los widgets Resultados o "Test & Score" de cada modelo para obtenerlos. El error cuadrático medio se encuentra descrito como "RSME" y el R Cuadrado Ajustado como R2 en Orange. La Figura 84 muestra cómo queda un archivo modificado en base a lo que se exportó de Orange.

Valores Re	Linear Reg	Neural Ne	Random Forest	Error RL	Error RN	Error RF		RSME RL	RSME RN	RSME RF
0	-0.00134	0.029146	0	0.001335	0.029146	0		1.246	1.358	1.349
0	0.147812	0.090637	0	0.147812	0.090637	0				
0	-0.01717	-0.00838	0	0.017174	0.008376	0				
0	-0.00837	0.012688	0	0.008369	0.012688	0				
0	0.011963	-0.13386	0	0.011963	0.133863	0		RCA RL	RCA RN	RCA RF
1	0.323257	0.329142	0.05	0.676743	0.670858	0.95		0.773	0.73	0.733
0	-0.12991	-1.03725	0	0.129913	1.037248	0				
0	-0.08292	-0.00261	0	0.082916	0.002609	0				
0	-0.06989	-0.12065	0	0.069893	0.120651	0				
0	-0.00474	0.103679	0	0.004738	0.103679	0				
0	-0.01154	0.006596	0	0.011536	0.006596	0		MODELO 2 -HOMBRES NO PROMOVIDOS OCTAVO BÁSICA		
0	0.017075	0.024079	0	0.017075	0.024079	0				

Figura 84. Archivo de extensión csv modificado con los errores correspondientes calculados y las demás métricas copiadas de los modelos de Orange. (Elaboración Propia)

Antes de pasar a comparar los resultados de cada modelo se precisan unas cuantas aclaraciones:

- El RSME o error cuadrático medio no tiene un "valor ideal", depende mucho de los datos con los que se está tratando.

- El valor obtenido en el RSME usualmente fluctúa debido a la cantidad de datos del repositorio. Cuando la cantidad de datos aumente, el RSME también, inclusive si es que no existe una peor predicción en el modelo.
- El valor obtenido del R2 o RCA sí tiene un "valor ideal", sin embargo, es irreal que se pretenda obtener el 100% de este valor debido a que no es posible crear un modelo perfecto. A partir de 65-70% puede considerarse un valor adecuado debido a la cantidad de datos con los que se está trabajando.
- Para complementar el análisis se hace uso del error absoluto y su media, valores que fueron calculados en cada uno de los archivos generados tanto con R como con Orange.
- Todos los modelos realizados con R se realizaron particionando la data tanto en un conjunto de entrenamiento como en un conjunto de pruebas, mientras que muchos de los modelos realizado con Orange fueron desarrollados usando el conjunto de datos original (los datos preprocesados provenientes del trabajo realizado en el Capítulo 3).
- Por el punto anterior, muchas veces no será posible realizar una comparación 1 a 1 debido a que las predicciones son arrojadas para un conjunto de aproximadamente 10000 filas (en el caso en el que se particiona la data) frente a predicciones en base a conjuntos de casi 30000 filas (en el caso en el que no se particiona la data).
- La inclusión de métodos adicionales sirve para mostrar cómo los RSME y RCA aumentan o disminuyen en cada modelo y así obtener mejores resultados, sin embargo, la inclusión de estos métodos adicionales no siguió una regla, es decir, en muchos modelos se incluyó.

5.3.1 Métricas Hombres Promovidos Segundo de Básica

Tabla 3- Capítulo 5
Métricas Hombres Promovidos Segundo de Básica – Versión 1

Regresión L	2010	2011	2012	2013	2014	2015	2016	2017
R2 RStudio	0.9360	0.9300	0.9230	0.9307	0.9381	0.9190	0.9373	0.9405
R2 Orange	0.9360	0.9300	0.9230	0.9307	0.9381	0.9190	0.9373	0.9405
RSME RStudio	3.5730	3.6013	3.3478	3.5249	3.7182	4.2634	4.2187	4.3120
RSME Orange	3.5730	3.6010	3.3480	3.5250	3.7180	4.2630	4.2190	4.3120
MAE RStudio	2.1128	1.9457	1.9721	2.0674	2.0884	2.4680	2.4819	2.5353
MAE Orange	2.1127	1.9457	1.9724	2.0674	2.0885	2.4679	2.4819	2.5353
Redes N.	2010	2011	2012	2013	2014	2015	2016	2017
R2	X	X	X	X	X	X	X	X
RSME	X	X	X	X	X	X	X	X
MAE	X	X	X	X	X	X	X	X

Métricas obtenidas con la variable Hombres Promovidos Segundo Básica Versión 1. (Elaboración Propia)

Tabla 4 - Capítulo 5
Métricas Hombres Promovidos Segundo Básica - Versión 2

Regresión L	2010	2011	2012	2013	2014	2015	2016	2017
R2 RStudio	0.9340	0.9300	0.9230	0.9306	0.9379	0.9186	0.9369	0.9450

R2 Orange	0 .9340	0 .9300	0 .9230	0 .9306	0 .9379	0 .9186	0 .9369	0 .9450
RSME RStudio	3 .5638	3 .6019	3 .3460	3 .5209	3 .7166	4 .2749	4 .2202	4 .3169
RSME Orange	4 .3050	3 .3640	3 .5470	3 .5470	3 .7060	4 .8640	4 .2210	4 .3050
MAE RStudio	2 .1022	1 .9439	1 .9663	2 .0569	2 .0819	2 .4541	2 .4738	2 .5362
MAE Orange	2 .4975	1 .9267	1 .9713	2 .0204	2 .0214	2 .4521	2 .4355	2 .4975
Redes N.	2 010	2 011	2 012	2 013	2 014	2 015	2 016	2 017
R2	0 .9220	0 .9130	X	X	X	X	X	0 .9220
RSME	4 .6910	3 .5900	X	X	X	X	X	4 .6910
MAE	2 .6944	2 .1116	X	X	X	X	X	2 .6944

Métricas obtenidas con la variable Hombres Promovidos Segundo Básica Versión 2. (Elaboración Propia)

Por lo tanto, se obtiene que:

- El R Cuadrado Ajustado (RCA/R2), como es de esperar, se mantiene igual en ambas herramientas.
- En general, el R Cuadrado Ajustado disminuye su valor del modelo 1 al modelo 2 en la aplicación con RStudio, mientras que ocurre lo contrario en el trabajo de Orange, en donde esta métrica aumenta levemente. Por lo tanto, la segunda construcción del modelo en Orange resulta mejor en este aspecto.
- Para el primer modelo, en ambos casos se trabajó separando los datos en conjuntos de entrenamiento y pruebas; precisamente aquí es donde no se nota una diferencia en el error cuadrático medio (RSME) entre RStudio y Orange.

- Para el Error Absoluto Medio (MAE) sucede lo mismo que con el RSME, los valores de ambos casos son prácticamente los mismos cuando se analizan las primeras versiones de los modelos.
- El RSME obtenido con los proyectos en Orange aumenta ligeramente para el modelo #2 de cada período lectivo, esto se debe a que en muchos de estos modelos se trabajó directamente con el archivo que contenía todos los datos luego del procesamiento, sin realizar la división. Vale recordar que en estos casos no se necesitó realizar la división de la data debido a la característica denominada "Validación Cruzada" que se puede activar en Orange y el software por sí solo se encarga de particionar la data. Sin embargo, el aumento no es considerable.
- El MAE se encuentra en un rango de entre 1.9 y 2.5, esto es la diferencia entre el valor original y el obtenido que no viene a representar un valor muy alto, por lo que se considera aceptable.
- Las métricas obtenidas usando Redes Neuronales no representan una mejora para los modelos en general, de hecho, muestran resultados similares, pero ligeramente inferiores a los obtenidos usando Regresión Lineal.

5.3.2 Métricas Mujeres Abandono Tercero de Bachillerato

Tabla 5 - Capítulo 5
Métricas Mujeres Abandono Tercero de Bachillerato – Versión 1

Regresión L	2010	2011	2012	2013	2014	2015	2016	2017
R2 RStudio	0.6587	0.6400	0.6989	0.7762	0.7320	0.7664	0.8428	0.9125
R2 Orange	0.6587	0.6400	0.6989	0.7762	0.7320	0.7664	0.8428	0.9125
RSME RStudio	0.3955	0.4315	0.4183	0.5506	0.5732	0.6210	0.5551	0.6336
RSME Orange	0.3960	0.4310	0.4180	0.5510	0.5730	0.6210	0.5550	0.9780

MAE RStudio	0 .0976	0 .1143	0 .1135	0 .1509	0 .1599	0 .1807	0 .1732	0 .1877
MAE Orange	0 .0976	0 .1144	0 .1136	0 .1510	0 .1599	0 .1806	0 .1732	0 .1877
Redes N.	2 010	2 011	2 012	2 013	2 014	2 015	2 016	2 017
R2	X	X	X	X	X	X	X	X
RSME	X	X	X	X	X	X	X	X
MAE	X	X	X	X	X	X	X	X

Métricas obtenidas con la variable Mujeres Abandono Tercero de Bachillerato Versión 1. (Elaboración Propia)

Tabla 6 - Capítulo 5
Métricas Mujeres Abandono Tercero de Bachillerato - Versión 2

Regresión L	2 010	2 011	2 012	2 013	2 014	2 015	2 016	2 017
R2 RStudio	0 .6579	0 .6400	0 .6988	0 .7759	0 .7319	0 .7663	0 .8425	0 .9112
R2 Orange	0 .6579	0 .6400	0 .6988	0 .7759	0 .7319	0 .7663	0 .8425	0 .9112
RSME RStudio	0 .3950	0 .4314	3 .3460	0 .5497	0 .5713	0 .6195	0 .5541	0 .6315
RSME Orange	0 .2150	0 .4490	3 .4560	0 .5160	0 .5710	0 .6460	0 .5280	0 .8230
MAE RStudio	0 .0911	0 .0969	1 .9663	0 .1438	0 .1550	0 .1712	0 .1652	0 .1793
MAE Orange	0 .0976	0 .1122	1 .9490	0 .1320	0 .1360	0 .1540	0 .1540	0 .1770
Redes N.	2 010	2 011	2 012	2 013	2 014	2 015	2 016	2 017
R2	X	0 .8970	X	0 .9170	0 .9340	0 .9350	0 .9410	0 .8230

RSME	X	0.2280	X	0.3010	0.2720	0.3290	0.3070	0.7740
MAE	X	0.0798	X	0.1040	0.0980	0.1660	0.1570	0.2460
Random F.	2010	2011	2012	2013	2014	2015	2016	2017
R2	X	0.8620	0.9910	0.9330	0.8740	0.9220	0.9430	X
RSME	X	0.2670	1.6980	0.2700	0.3760	0.3610	0.3040	X
MAE	X	0.0375	0.8140	0.047	0.0500	0.0540	0.0520	X

Métricas obtenidas con la variable Mujeres Abandono Tercero de Bachillerato Versión 2. (Elaboración Propia)

Por lo tanto, se obtiene que:

- Para el caso del R2/RCA en el primer modelo realizado en ambas herramientas, se muestra un incremento constante con el paso de los períodos lectivos, inclusive se termina con un RCA para el último año lectivo de más del 90%.
- Para los primeros modelos, entre el trabajo con RStudio y Orange existen diferencias mínimas para el RSME y el MAE.
- Comparando el primer modelo con el segundo de cada período lectivo, se puede notar como el RCA a penas cambia, esto es debido a que en la construcción solamente varió el grupo de variables elegido, más no el proceso en sí. Lo que muestran estos valores es que, efectivamente, las variables descartadas no tenían ninguna clase de repercusión en los modelos.
- Dado que las métricas que se obtenían en ambas herramientas no estaban variando, se presenta la inclusión de uno o dos métodos adicionales (Redes Neuronales y/o Random Forest). Con excepción del primero, todos los períodos lectivos fueron sometidos a estos métodos adicionales para obtener nuevos resultados. Para mejorar las métricas también se emplea un modo distinto de seccionar los datos en Orange.
- Los resultados con Redes Neuronales y Random Forest son considerablemente mejores que los obtenidos con Regresión Lineal, destacando el MAE obtenido con Random Forest en casi todos los períodos lectivos. Con lo que se concluye

que el cambio en la forma de particionar los datos en Orange tuvo efectos positivos para la construcción de los modelos.

- Comparando los datos obtenidos con Redes Neuronales y Random Forest se encuentra que en algunos períodos lectivos usar Redes Neuronales es mejor, mientras que en otro usar Random Forest es mejor, pero siempre presentando diferencias mínimas y valores altos, por lo que ambos métodos son recomendables de usar en este caso.

5.3.3 Métricas Hombres No Promovidos Octavo de Básica

Tabla 7 - Capítulo 5
Métricas Hombres No Promovidos Octavo de Básica - Versión 1

Regresión L	2010	2011	2012	2013	2014	2015	2016	2017
R2 RStudio	0.8465	0.7153	0.7631	0.5462	0.7758	0.8182	0.7838	0.8276
R2 Orange	0.8465	0.7153	0.7631	0.5462	0.7758	0.8182	0.7838	0.8276
RSME RStudio	2.7316	1.5025	1.2955	1.0215	0.9121	1.0147	1.0784	1.1821
RSME Orange	2.9970	1.5550	1.3540	1.2150	1.0460	1.1340	1.1770	1.2480
MAE RStudio	0.6478	0.3741	0.3440	0.2746	0.2794	0.3463	0.3872	0.4079
MAE Orange	0.5720	0.3790	0.3300	0.2650	0.2650	0.3490	0.3830	0.4030
Redes N.	2010	2011	2012	2013	2014	2015	2016	2017
R2	0.7330	X	0.6090	X	X	0.7170	X	X
RSME	2.7020	X	1.3690	X	X	1.2370	X	X
MAE	0.6210	X	0.4160	X	X	0.4530	X	X

Random F.	2010	2011	2012	2013	2014	2015	2016	2017
R2	0.6150	X	0.6110	X	X	X	X	X
RSME	3.2470	X	1.3640	X	X	X	X	X
MAE	0.3900	X	0.2770	X	X	X	X	X

Métricas obtenidas con la variable Hombres No Promovidos Octavo de Básica Versión 1. (Elaboración Propia)

Tabla 8 - Capítulo 5
Métricas Hombres No Promovidos Octavo de Básica - Versión 2

Regresión L	2010	2011	2012	2013*	2014	2015	2016	2017
R2 RStudio	0.8462	0.7151	0.7626	0.5446	0.7715	0.8151	0.7786	0.8241
R2 Orange	0.8462	0.7151	0.7626	0.5446	0.7715	0.8151	0.7786	0.8241
RSME RStudio	2.6890	1.5016	1.2942	1.0069	0.8995	1.0160	1.0694	1.1838
RSME Orange	2.6860	1.3820	5.8270	1.1520	1.0110	1.1220	1.1170	1.1930
MAE RStudio	0.6252	0.3646	0.3287	0.2579	0.2482	0.3275	0.3569	0.3786
MAE Orange	0.5190	0.3560	1.9090	0.2420	0.2330	0.3300	0.3600	0.3950
Redes N.	2010	2011	2012	2013	2014	2015	2016	2017
R2	0.7340	0.9230	0.9350	0.9300	0.9400	0.6910	0.9290	0.9290
RSME	2.7000	0.6560	5.7840	0.4590	0.4690	1.2940	0.5980	0.6970
MAE	0.6390	0.2820	1.9840	0.2120	0.2050	0.4550	0.2880	0.2960

Random F.	Ra 2010	2 011	2 012	2 013	2 014	2 015	2 016	2 017
R2	0 .6170	0 .9020	0 .9340	0 .8820	0 .9180	0 .7160	0 .9330	0 .9350
RSME	3 .2360	0 .7440	5 .8410	0 .5940	0 .5490	1 .2390	0 .5810	0 .6650
MAE	0 .3860	0 .1180	1 .4440	0 .0910	0 .0950	0 .2710	0 .1340	0 .1520

Métricas obtenidas con la variable Hombres No Promovidos Octavo de Básica Versión 2. (Elaboración Propia)

Por lo tanto, se obtiene que:

- En este caso en particular, no se muestra un crecimiento constante del RCA, sino más bien éste se mantiene fluctuando en alrededor del 80%, tanto para las primeras como segundas versiones de los modelos, pero se presenta una excepción en el año 2013 en donde se obtiene un valor muy bajo. En contraste, el RSME y el MAE presentan valores bajos que podrían proceder de un modelo con un mejor RCA. Por lo que, el modelo de este período lectivo de todas formas se considera aceptable.
- Para los casos de los métodos adicionales aplicados en las primeras versiones de los modelos, se usó la misma forma de particionar los datos, por lo que como se muestra en la tabla, no existe una mejora sustancial en estos valores.
- Se mantiene la pequeña diferencia del RSME y el MAE entre los trabajos desarrollados con RStudio y Orange, para las dos versiones de cada modelo, al igual que sucedía con los modelos de las otras variables analizadas.
- Para el resto de períodos lectivos, los métodos de Redes Neuronales y Random Forest arrojan mejores métricas que las correspondientes a Regresión Lineal. Debido a la reiteración de resultados, esta mejora se le atribuye al cambio en la forma en que se particionan los datos para poder trabajar, así como un mejor ajuste de estos métodos al conjunto de datos con el que se trabaja.
- El período lectivo 2013 (2012-2013) tiene un asterisco puesto que también fueron aplicados los métodos kNN y AdaBoost, dos métodos reconocidos en el ámbito de Aprendizaje Automático, sin embargo, presentaron resultados pobres con kNN

teniendo un RCA de 0.4 y AdaBoost de 1.00. El primer valor es muy bajo como para ser considerado aceptable y el segundo de 1.00 es técnicamente imposible puesto que no existe un método que se ajuste a la perfección al conjunto de datos.

- En contraste con las primeras versiones de los modelos, en todas las segundas versiones fueron aplicados los métodos de Redes Neuronales y Random Forest. Un caso interesante por resaltar es que para el período lectivo 2011-2012 los trabajos con Orange presentan un considerable aumento de RSME y MAE con respecto a sus contrapartes de RStudio. Debido a esto, se puede concluir que los modelos de Orange no son óptimos y se consideraría usar el modelo de RStudio en esta situación en particular.

5.4 Cuadro Comparativo de Herramientas

Tabla 9
Cuadro Comparativo de Herramientas

	Paradigma de Programación	Métodos Disponibles para construcción de Modelos.	Gráficos disponibles para visualización de datos.	Bibliotecas	Trabajo con otros lenguajes de programación	Manejo de Datos	Tiempos de Respuesta
R/RStudio	Programación Funcional	Limitado a la capacidad del programador.	Limitado a las bibliotecas usadas.	Las que estén disponibles para agregar.	Integración con varios lenguajes como Python o Java.	Sin límites para tamaño de archivos. Soporta grandes volúmenes de datos.	5.31 segundos para obtener resultados de un modelo con Regresión Lineal aprox.
Orange	Programación Visual	20 métodos disponibles (aproximadamente en la versión usada).	17 distintos gráficos (aproximadamente en la versión usada)	No se pueden agregar bibliotecas.	Trabajo únicamente con Python.	Soporte parcial para grandes volúmenes de datos; sin límite de tamaño, pero problemas visuales con archivos con muchos datos.	10.12 segundos (Regresión Lineal) aprox. 45.97 segundos (Redes Neuronales) aprox.

Cuadro comparativo, mostrando las principales características de las dos herramientas. (Elaboración Propia)

6. Conclusiones y Recomendaciones

6.1 Determinación de ventajas y desventajas de las herramientas seleccionadas.

Se desglosan las ventajas y desventajas de cada herramienta partiendo del cuadro comparativo a modo de resumen mostrado en el capítulo 5.

6.1.1 Ventajas del trabajo en RStudio

El trabajo con RStudio se basó en la utilización de tres bibliotecas de libre acceso las cuales sirvieron para completar un script general usado para cada grupo de datos de los períodos lectivos analizados. Los resultados arrojados por estos modelos se especifican en los archivos correspondientes, así como en la recopilación realizada en el capítulo 5. La construcción misma de los modelos y el análisis comparativo realizado previamente llevan a las siguientes conclusiones en forma de ventajas:

- No es necesario aprender el lenguaje de programación R para construir modelos de Aprendizaje Automático o Ciencia de Datos, con aprender unas cuantas funciones y el manejo de la interfaz principal del IDE que se esté utilizando (en este caso RStudio) será suficiente para poder analizar los datos correspondientes. El hecho de haber partido de un proceso general de Minería de Datos facilita la construcción de los modelos.
- El manejo de RStudio es muy similar a otros software enfocados en el análisis de datos como Anaconda Python, por lo que su uso se vuelve sencillo al haber tenido experiencias similares durante el transcurso de la carrera.
- El hecho de tener un gran número de bibliotecas expande las posibilidades del software, de modo que se pueden simplificar procesos por medio de funciones importadas. En este caso en particular se pudo importar una biblioteca que permitiera la elaboración de gráficos más avanzados para analizar los datos, así como las mismas funciones de construcción de los modelos para la Minería de Datos.
- Importar y exportar datos en R se puede solucionar por medio de un corto comando y se puede preparar los datos para que cuando se creen los archivos exportados, éstos tengan el formato y la apariencia deseadas.

- Cuando se trabaja directamente con un lenguaje de programación, las limitantes son impuestas en mayor medida por la capacidad del programador que del lenguaje mismo. Trabajar en la construcción de modelos escribiendo el código, línea por línea, da un control total sobre lo que se está realizando. Muchas carencias de R se ven resueltas con las facilidades para incorporar segmentos de código de otros lenguajes, que pueden aportar en algunas secciones del modelo para llegar a un mejor resultado.

6.1.2 Desventajas del trabajo en RStudio

Asimismo, fueron encontradas algunas desventajas de trabajar con este lenguaje de programación en el IDE RStudio:

- Si bien es totalmente cierto que no es necesario tener un conocimiento profundo del lenguaje de programación para construir esta clase de modelos predictivos, cuando se importan bibliotecas, éstas usualmente vienen con lo que se denomina "gramática" que viene a ser el control específico de cada función relacionada con dicha biblioteca. Como existen muchos autores para las numerosas bibliotecas disponibles, el manejo de las funciones no siempre se realiza de la misma manera, lo que puede ser engorroso para el programador.
- Cuando se presenta el resumen de cada modelo generado, éste no cuenta con todas las métricas que uno podría esperar, teniendo que calcular manualmente algunas de ellas como, por ejemplo, el RSME o el MAE.
- Muchas personas ajenas al mundo de la programación podrían encontrar un tanto problemático el hecho de tener que escribir un comando o función para obtener un resultado. Si bien se vuelve a hacer hincapié en el hecho de que no se necesita conocer mucho para crear esta clase de modelos, algunos profesionales prefieren lo más sencillo para completar un trabajo, puesto que esta clase de análisis es solo una parte de sus actividades.

6.1.3 Ventajas del trabajo en Orange

El trabajo en Orange se basó en el aprendizaje de este software para realizar un modelo sencillo que pueda cumplir los mismos objetivos planteados alcanzados con RStudio. De este trabajo se obtiene las siguientes conclusiones a modo de ventajas que se listan a continuación:

- El trabajo con Orange es enteramente visual; en este software no existe la necesidad de introducir líneas de código para obtener alguna respuesta, sino que todo se realiza de manera gráfica con la concatenación de esferas con algún propósito en específico, que el programa denomina “widgets”. Esto produce que se cree una especie de grafo por cada modelo realizado, donde el usuario puede ir viendo de mejor manera lo que el modelo está realizando para lograr el resultado.
- El manejo del software se produce casi de manera natural, donde una breve explicación será suficiente para que una persona, con objetivos claramente definidos, use este programa y alcance resultados satisfactorios en poco tiempo.
- Orange es una herramienta enfocada totalmente en el Aprendizaje de Máquina. Todas las opciones que este software provee a los usuarios están pensadas para Minería de Datos, Análisis de Datos y demás propósitos similares. La gestión de los datos es su principal función.
- Orange es una herramienta que encaja perfectamente en la construcción de modelos predictivos, como los presentados en este trabajo; desde el preprocesamiento o visualización, hasta la forma en que se presentan los resultados de los modelos, incluyendo un resumen de métricas para que el usuario tenga una idea de qué tan bien ha ido su trabajo.
- Orange no tiene ninguna clase de restricción con el usuario en cuanto a la monetización, puesto que es un software open-source y 100% gratuito, esto es, no cuenta con versiones de pago con mejores prestaciones, una modalidad fácil de encontrar en la industria y que termina segregando a la comunidad, obligando a los usuarios a saltarse a la versión de pago cuando las características de la versión gratuita no les sean suficientes.
- Una de las principales ventajas de este software, es que puede ser usado como una librería de Pyhton, así como también permite cargar scripts escritos en este lenguajes de programación para ampliar sus opciones de manera considerable.
- Orange tiene algunas opciones para particionar los datos de manera automática, en caso de ser necesario (para los métodos de Aprendizaje Supervisado). Se puede usar la opción de “Validación Cruzada” para que el programa mismo sea el que divida los datos en base a unos parámetros establecidos; también se puede trabajar

independientemente (como se lo hizo en algunos modelos presentados) para que Orange trabaje sobre los datos de entrenamiento y luego pruebe el modelo sobre los datos de prueba. Esto, se puede realizar en RStudio, pero se necesitaría de procesos más largos que implican más líneas de código; con Orange es cuestión de activar o desactivar un radio button.

6.1.4 Desventajas del trabajo en Orange

De la misma manera que se hallaron ventajas, se muestran las desventajas de trabajar con Orange a continuación:

- El hecho de poder incluir scripts de Python limita la naturaleza de este software (la programación visual) y esto implica que las opciones de Orange sean reducidas a lo que este programa ofrece. Ya sea que con el paso del tiempo se inventen nuevos métodos para aplicar Aprendizaje de Máquina o que se mejoren los ya establecidos, el hecho de que Orange cuente con un conjunto de métodos predefinidos deja al usuario sin la oportunidad de aprovechar esta herramienta con los mencionados nuevos o mejorados métodos hasta que los desarrolladores de Orange decidan incorporarlos en forma de nuevas versiones del software.
- Orange tiene unos cuantos problemas relacionados con la carga de datos. Cuando la cantidad de datos es muy grande, tal y como lo son los archivos preprocesados .csv de cada período lectivo, el programa tiene que cargar los datos cada vez que se inicie. Por suerte, Orange deja la última fuente usada de la computadora para que el usuario solo tenga que abrir nuevamente el archivo y éste sea cargado. Esto implica que cada vez que el archivo tenga que ser cargado, el modelo va a ser ejecutado nuevamente, arrojando nuevas predicciones, aunque sin ninguna variación considerable en las métricas si el modelo no fue modificado en ninguna sección.
- Es posible que el usuario experimente en algunas ocasiones una demora considerable hasta que Orange procese lo que se está solicitando. Esto pasa generalmente cuando se construye un modelo y se añaden varios métodos para obtener resultados.
- El problema de tener que cargar los archivos de datos cada vez que se inicie Orange, sumado a los minutos que se debe esperar en algunas ocasiones para que

el software vuelva a calcular los resultados y la imposibilidad de ver los resultados previos mientras no Orange no haya terminado de procesar los modelos, puede llevar a una frustración por parte del usuario.

- Existen pocas herramientas de etiquetado en el programa, que no van más allá de incluir cuadros de texto y flechas. Los flujos de trabajo usualmente quedan desorganizados debido a la falta de opciones que tiene el software en este sentido. Tampoco presenta cuadros de advertencia en caso de existir errores en la concatenación de widgets o problemas con las entradas/salidas respectivas, lo que se compensa con un pequeño ícono encima del widget con problemas. Por supuesto, esto no es suficiente para que el usuario comprenda lo que salió mal, aunque no termina impidiendo la construcción general de un modelo.

Una vez se concluye el trabajo se entiende la importancia de trabajar con dos herramientas tan distintas, aunque diseñadas para cumplir un objetivo en común. Es clara la diferencia de trabajar con una u otra herramienta, teniendo un método tradicional con RStudio, que es el de programar y tener un control amplio sobre lo que la máquina debe realizar; mientras que Orange supone una forma sencilla de construir modelos de manera visual, solamente buscando el widget que sea necesario y unirlo con otros para llegar a un resultado determinado. A continuación, se presentan algunas conclusiones sobre el trabajo realizado:

- Una compañía que esté interesada en trabajar con sus datos, realizando modelos predictivos de Minería de Datos, debería buscar mantener un control estricto sobre los mismos, por lo que trabajar con un software de manera tradicional como lo es el caso de RStudio será su mejor alternativa.
- Orange es una herramienta completamente funcional que provee resultados muy similares a lo visto con RStudio. Si bien es cierto tiene sus restricciones (en especial para las personas que no desean incluir código de Python), muchas de las principales maneras de construir un modelo están presentes.
- Se necesita complementar el análisis per se con la visualización de los datos, tal y como se demostró en una de las comparativas realizadas en el capítulo 5. Muchas veces, los gráficos nos pueden mostrar cómo se está comportando la data y de esta se tiene una prueba adicional para lo logrado con las métricas y los modelos.

- Todo el trabajo realizado se basa en la construcción de modelos predictivos, debido a la forma en que se encontraban los datos fuentes, con muchos datos numéricos y pocos datos tipo texto relevantes. La aplicación de Minería de Datos sobre un conjunto de datos no necesariamente implica descubrir patrones, también se pueden predecir valores y obtener una ventaja competitiva para una organización. Hablando de este mismo ejemplo, el de las instituciones de educación en el país, se podrían predecir valores para tener claro cómo va a evolucionar un sector determinado de estudiantes, o descubrir que en la Costa se deben aplicar más controles para evitar el abandono de estudios en ciertos años del Bachillerato, entre otros ejemplos. La gente suele tener la idea equivocada de que aplicar Minería de Datos es encontrar mágicamente información sin ningún proceso previo, como aplastar un botón y que algún programa muestre no datos sino información precisa y procesada. Es evidente que así no funcionan las cosas y por eso es por lo que se llevan a cabo estos trabajos para orientar un poco a la gente que no tenga mucha idea del tema.
- Los conjuntos de datos obtenidos del Ministerio de Educación se encontraron relativamente organizados, con pocos datos vacíos en las numerosas columnas de cada archivo de Excel. Esto facilitó en gran medida el trabajo, sin embargo, este escenario ideal no siempre se produce, es por lo que se necesita dar la importancia que tiene al preprocesamiento de datos, el cual es necesario y puede llegar a implicar una gran serie de operaciones hasta dejar al conjunto de datos elegido en condiciones favorables para los análisis.

6.2 Ética en la Minería de Datos

Como se mencionó previamente, la Minería de Datos no solo busca encontrar patrones, sino también generar predicciones en base a la construcción de modelos que permitan esto. Ya sean patrones o predicciones, todo lo que se consigue con la Minería de Datos debe ser parte de un proceso consensuado con los respectivos dueños de los datos.

En un mundo en donde existen compañías que basan sus negocios en bienes intangibles como lo son sus datos e información, no se puede permitir que cualquiera tenga derecho a analizar dichos bienes, mucho menos sabiendo que existen herramientas gratuitas, disponibles para todos, con las cuales se pueden generar ventajas competitivas.

El proceso de Minería de Datos puede acabar sin la construcción de información, es decir, se puede usar una herramienta para obtener patrones o predicciones y nada más. A pesar de que sería óptimo que la misma persona que realizó todo el proceso, sea la que termine por concluir los análisis, esto no siempre pasa y depende de cada persona dueña de los derechos de los datos el permitir o no que los mismos sean analizados. La ética de una persona consiste en llevar a cabo su trabajo y cumplir con lo que se le estipuló, no ir más allá por el hecho de tener la posibilidad de afectar a la compañía en la que se desempeña como profesional, siguiendo una línea de principios para no sobrepasar funciones, teniendo control sobre las acciones que se pueden hacer y firmeza para no cometer actos que atenten contra otras instituciones; estos son los principios que fueron inculcados como parte de la enseñanza general en la carrera de Ingeniería en Sistemas y Computación en la Pontificia Universidad Católica del Ecuador

Cuando se habla de ventajas competitivas, realmente se tiene que pensar en casos en los que un simple análisis de datos puede llevar a rivales a sacar provecho de ciertos sectores de una empresa que no se encuentren bien o generar predicciones para ver si un líder del mercado va a seguir vendiendo bien un producto en el futuro próximo.

Este trabajo prueba que es factible construir modelos eficientes de manera sencilla para poder generar predicciones y, de manera secundaria, encontrar relaciones entre variables incluidas en los modelos para llegar a construir información; si a lo anterior se le suma el análisis que se puede realizar con distintos gráficos generados a partir de los datos, entonces se determina que existe un riesgo para todos los que tengan datos comprometedores. Por esto, en una compañía se deben establecer una serie de acciones que parten desde la generación de contratos con los profesionales relacionados para que éstos tengan restringido el derecho de analizar datos sin el previo consentimiento de las autoridades supervisoras.

Hoy en día, se vuelven más populares las actividades derivadas de la Inteligencia Artificial, como el Aprendizaje de Máquina o la Minería de Datos debido a la gran cantidad de aplicaciones que se pueden encontrar, esto implica un control mayor por parte de todos aquellos que tengan en riesgo datos personales o de alguna entidad privada, pero también implica que los profesionales capaces de llevar a cabo estos análisis de los datos sean conscientes de lo que se puede llegar a generar, para que todo lo que se haga, se lo realice de manera responsable.

6.2 Conclusiones de la aplicación de Minería de Datos sobre los datos del Ministerio de Educación

Las conclusiones respectivas del trabajo realizado se muestran a continuación:

- La fuente de datos elegida resultó ser lo suficientemente confiable como para poder trabajar sin mucha dificultad, teniendo los datos organizados en archivos .xlsx (Excel) separados por período lectivo, lo cual facilitó el preprocesamiento de datos.
- El proceso de Minería de Datos fue cumplido con éxito, siguiendo todas las etapas propuestas en la teoría, las cuales fueron consecutivamente exploradas durante el estudio; dentro de estas etapas, el preprocesamiento resultó ser parte clave para poder obtener los resultados esperados ya que simplificó el trabajo posterior.
- Las herramientas elegidas, en líneas generales, no presentaron inconvenientes a la hora de usar su interfaz, sin embargo, sí se pueden notar algunos problemas con el tema de la carga de datos o el despliegue de gráficos teniendo en cuenta la cantidad de variables de cada uno de los archivos del Ministerio de Educación.
- Los datos elegidos resultan ser en su mayoría valores numéricos en forma de los datos de estudiantes de cada nivel de educación desde el Primero de Básica hasta el Tercero de Bachillerato, lo cual llevó a la elección de un análisis predictivo, dejando de lado la extracción de patrones por la falta de variables cualitativas dentro de los archivos.
- La delimitación del número de períodos lectivos fue clave para continuar con el proceso de modo que éste no se extendiera de más, debido a que se crearon varios modelos para cada uno de los períodos, con lo que elevar el número de años analizados hubiera alargado de más el estudio.
- A pesar de los beneficios descritos del conjunto de archivos, también es necesario mencionar que los mismo deberían estar en una base de datos relacional o un data Warehouse de modo que se pueda sacar más provecho a los datos.

6.3 Recomendaciones para futuras aplicaciones de Minería de Datos sobre fuentes similares

El trabajo llevado a cabo con los datos elegidos, también ha dejado algunas recomendaciones para las personas interesadas en replicarlo. Estas recomendaciones se listan a continuación:

- El trabajo debe empezar teniendo claro lo que se va a realizar. Existen muchas formas de aplicar Minería de Datos, partiendo principalmente de si se va a realizar un análisis predictivo o si se van a extraer patrones de los datos. Aunque se pueden complementar, estas dos aproximaciones tienen un desarrollo distinto y se debe partir con objetivos claros.
- La fuente de los datos debe ser analizada exhaustivamente. Muchas veces los datos no se encuentran bien formateados, ni siquiera en la misma extensión de archivo. El preprocesamiento es parte fundamental del proceso y debe ajustarse tanto para tener una misma forma de datos en todos los archivos involucrados, como para poder ser utilizados en las herramientas que se elijan.
- Inclusive teniendo las herramientas elegidas y el tipo de análisis que se quiere realizar, no existe un método único para llevar a cabo la Minería de Datos; se recomienda, por lo tanto, encontrar la manera más simple de construir modelos, sin que esto afecte a la eficacia de los mismos.
- Trabajar con Orange tiene sus ventajas y desventajas, como ya fue mencionado previamente; se recomienda exportar los archivos cada vez que se genere un modelo ya que, al cerrar y abrir el programa, es posible que las fuentes de datos tengan que ser cargadas nuevamente y las predicciones conseguidas sean alteradas ligeramente.
- Es importante seguir los pasos descritos en este trabajo para completar el proceso de Minería de Datos con el fin de tener modelos organizados y entendibles en caso de que alguien más desee trabajar sobre los mismos.
- Si durante la carrera se trabajó con algún software para análisis de datos, como en este caso se tenía experiencia previa con Anaconda Python, elegir una herramienta como RStudio es lo más recomendable, puesto que la construcción de

scripts puede compartir lógica con lo que se realizó en otras herramientas de este estilo.

- La carrera provee a los estudiantes las facilidades necesarias para poder realizar un trabajo de este estilo, en forma de materias importantes como Inteligencia Artificial o Aplicaciones Difusas y en general las materias relacionadas con programación. Durante el transcurso de la carrera, los estudiantes adquieren experiencia para que al trabajar con software de análisis de datos no se tengan dificultades de ningún tipo. Aunque es evidente que la experiencia viene del trabajo en circunstancias cotidianas que se dan en la vida real, como la aplicación de algún análisis de este tipo para poder afianzar ciertos datos de una organización.
- Comparar herramientas implica un trabajo extenso si se decide usar un amplio conjunto de datos, como es este caso, por lo que, reducir el número de herramientas a dos o tres es algo fundamental para no alargar el trabajo innecesariamente, con especial énfasis en la elección correcta de herramientas para poder tener dos maneras de trabajo distintas que puedan ser comparadas para demostrar sus diferencias.
- No se debe apresurar los análisis, sino más bien realizarlos uno por uno hasta tener todos los datos suficientes para su respectiva comparación.

7. Bibliografía

Casas, J., Gironés, J., Minguillón, J., & Caihuelas, R. (2017). *Minería de Datos: Modelos y Algoritmos.* Barcelona: Editorial UOC.

Coppola, C., & Neelley, E. (2004). Open source - opens learning: Why open source makes sense for education. *rSmart*, 1-3.

Demšar, J., & Zupan, B. (12 de Noviembre de 2012). Orange: Data Mining Fruitful and Fun - A Historical Perspective. Liubliana, Eslovenia.

Ghahramani, Z. (2004). Unsupervised Learning. En Z. Ghahramani, *Unsupervised Learning* (págs. 77-112). Berlín: Springer.

Gibert, K., Ruiz, R., & José, R. (2006). Minería de Datos: Conceptos y Tendencias. *Revista Iberoamericana de Inteligencia Artificial*, 11-18.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. En J. Han, M. Kamber, & J. Pei, *Data Mining: Concepts and Techniques* (págs. 1-8). Waltham: Elsevier Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning. En T. Hastie, R. Tibshirani, & J. Friedman, *The Elements of Statistical Learning* (págs. 43-92; 485-520). Stanford: Springer.

IBM. (2012). Manual CRISP-DM de IBM SPSS. Nueva York.

International Journal of Engineering Development and Research. (2017). Analysis of Data Using Data Mining tool. *International Journal of Engineering Development and Research*.

Kabacoff, R. (2017). *Statmethods*. Obtenido de Graphics with ggplot2: https://www.statmethods.net/advgraphs/ggplot2.html

Kukasvadiya, M., & Divecha, N. (2017). Analysis of Data Using Data Mining tool. *International Journal of Engineering Development and Research*.

Olson, D., & Delen, D. (2008). Advanced Data Mining Techniques. En D. Olson, & D. Delen, *Advanced Data Mining Techniques* (págs. 9-25). Lincoln: Springer.

Palma, C., Palma, W., & Pérez, R. (2009). *Data Mining: El arte de Anticipar.* Santiago de Chile: RIL Editores.

R Core Team. (2018). *Cran.r Project*. Obtenido de Cran.r Project: https://cran.r-project.org/doc/manuals/r-release/R-lang.pdf

RStudio. (2018). *RStudio*. Obtenido de RStudio - About Us: https://www.rstudio.com/about/

Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. En S. Russell, & P. Norvig, *Artificial Intelligence: A Modern Approach* (3rd ed., págs. 693-854). New York: Pearson Prentice Hall: Upper Sadlle River.

Suresh, S., Sundararajan , N., & Savitha, R. (2012). Supervised Learning with Complex-valued Neural Networks. En S. Suresh, N. Sundararajan, & R. Savitha, *Supervised Learning with Complex-valued Neural Networks* (págs. 1-29). Springer.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Pearson.

Tufféry, S. (2011). Data Mining and Statistics. En S. Tufféry, *Data Mining and Statistics* (págs. 10-15). Rennes: Wiley.

University of Ljubljana. (2019). *Orange BioLab*. Obtenido de Orange Docs: https://orange.biolab.si/docs/

Zhao, Y. (2012 de Diciembre de 2012). R and Data Mining: Examples and Case Studies. (Elsevier, Ed.)