PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR
MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

ANÁLISIS PREDICTIVO DE LA DEMANDA DE ESPECIES VALORADAS EN EL CONSULADO DEL ECUADOR EN QUEENS, ESTADOS UNIDOS DE AMÉRICA

Trabajo de titulación previo a la obtención del título de Magíster en Sistemas de Información mención Data Science

Línea de investigación: Adquisición, preprocesamiento, gestión y gobernanza de datos

Autor:
Victor Alfonso Chiza Monarco

Tutor del trabajo de titulación
Dr. Rafael Melgarejo

Quito – Ecuador
Agosto, 2024

ANÁLISIS PREDICTIVO DE LA DEMANDA DE ESPECIES VALORADAS EN EL CONSULADO DEL ECUADOR EN QUEENS, ESTADOS UNIDOS DE AMÉRICA

Contenido

Índice de figuras

Índice de tablas

1. Capítulo I: Introducción

1.1 Antecedentes

En el Ecuador se ha observado una escasez de investigaciones específicas que aborden de manera integral y aplicada la gestión de los trámites consulares y se ha identificado la necesidad de mejorar la gestión del stock de especies valoradas utilizadas en los trámites en el Consulado del Ecuador en Queens. Aunque se han realizado algunos análisis generales sobre la gestión consular, se ha observado una falta de estudios específicos que aborden de manera detallada el análisis y la predicción de la demanda de estas especies valoradas.

La gestión del stock de especies valoradas es un aspecto crítico para garantizar la eficiencia en la prestación de trámites consulares. Sin embargo, la falta de una metodología precisa para estimar la demanda futura de estas especies valoradas dificulta la planificación estratégica y puede llevar a una asignación inadecuada de recursos.

En este sentido, el (Ministerio de Relaciones Exteriores y Movilidad, 2020) proporciona información relevante sobre el arancel consular y diplomático vigente, donde se define una especie valorada como cualquier elemento utilizado para respaldar el cobro de tasas por las actuaciones previstas en el Arancel, autorizadas por el Ministerio de Finanzas. Esto incluye libretines de pasaporte, formularios, estampillas, timbres, sellos, entre otros. Esta definición es crucial para comprender la importancia y el uso de las especies valoradas en los trámites consulares y la necesidad de una gestión eficiente de estos recursos.

1.2 Planteamiento del problema

El Consulado del Ecuador en Queens brinda una gran variedad de trámites consulares o también llamadas actuaciones, abarcando desde actos notariales y poderes hasta trámites de registro civil y renovación de pasaportes. Para llevar a cabo estos trámites, se requiere el uso de especies valoradas, que son elementos utilizados para respaldar los cobros realizados por los trámites consulares. Estas especies valoradas tienen diferentes denominaciones y valores, y su gestión eficiente es esencial para garantizar un servicio consular eficaz y oportuno.

En la actualidad, la proyección mensual de la demanda de las especies valoradas se basa en estimaciones que pueden no reflejar con precisión las necesidades reales. El encargado de la bodega de especies valoradas debe revisar mensualmente el stock y proyectar mensualmente qué especies han sido utilizadas en mayor cantidad y necesitan ser reabastecidas. Además, debe considerar cuidadosamente qué denominaciones aún están disponibles en la bodega y pueden ser utilizadas antes de solicitar nuevas especies, ya que, dependiendo del valor de los trámites consulares, se pueden asignar especies valoradas de diferente denominación hasta completar el monto requerido.

Esta falta de comprensión detallada de la demanda de cada tipo de trámite y el uso asociado de especies valoradas dificulta la planificación estratégica y la asignación eficiente de recursos en el Consulado del Ecuador en Queens.

En este contexto, surge la necesidad de implementar un análisis predictivo que permita anticipar la demanda futura de especies valoradas con mayor precisión. Al emplear técnicas avanzadas de ciencia de datos, como el análisis de datos históricos de trámites consulares y el uso de especies valoradas, se pueden desarrollar modelos predictivos que mejoren la precisión de las proyecciones de demanda y optimicen la gestión del stock de especies valoradas en el Consulado.

1.3 Justificación

La optimización de la gestión del stock de especies valoradas y la eficiente asignación de este tipo de recursos en el Consulado del Ecuador en Queens son fundamentales para garantizar un servicio consular ágil y efectivo a los ciudadanos ecuatorianos en Estados Unidos. Dada la variedad de trámites consulares y la necesidad de contar con las especies valoradas adecuadas para cada uno, es muy importante comprender y prever la demanda de estos recursos de manera precisa.

El análisis y la predicción de la demanda de especies valoradas permitirán al Consulado anticipar las necesidades futuras de trámites consulares y planificar apropiadamente el abastecimiento de especies valoradas en su stock. Esto no solo mejorará la eficiencia en la prestación de servicios consulares, sino que también evitará la escasez o el exceso de estos insumos, optimizando así el uso de recursos financieros y logísticos.

Al implementar este proyecto, se espera mejorar significativamente la calidad y la eficiencia de los servicios consulares ofrecidos por el Consulado del Ecuador en Queens, lo que contribuirá a

fortalecer la relación entre la comunidad ecuatoriana residente en Estados Unidos y las autoridades consulares. Además, esta iniciativa ayudará a promover una gestión más transparente y eficaz de los recursos públicos asignados a los servicios consulares.

1.4 Objetivos

1.4.1 Objetivo general

Predecir la demanda de especies valoradas utilizadas en los trámites consulares esperados en el Consulado del Ecuador en Queens, Estados Unidos de América, con el fin de optimizar la gestión del stock y mejorar la eficiencia en la asignación de este tipo de recursos.

1.4.2 Objetivos específicos

- Identificar los tipos de trámites consulares más frecuentes y su correspondiente demanda de especies valoradas en el Consulado del Ecuador en Queens, con el fin de facilitar el análisis de patrones de demanda.
- Analizar los patrones históricos de uso de especies valoradas de los años 2021, 2022 y 2023 para cada tipo de trámite consular en el Consulado del Ecuador en Queens, con el propósito de comprender mejor la distribución y el comportamiento de los datos históricos para la construcción de modelos predictivos.
- Desarrollar modelos predictivos precisos para estimar la demanda futura de especies valoradas para los trámites consulares esperados, considerando las tendencias históricas y otros factores relevantes, con el objetivo de mejorar la planificación estratégica y la gestión del stock de especies valoradas.
- Evaluar la efectividad de los modelos predictivos y ajustar parámetros para mejorar su capacidad predictiva, garantizando así la precisión y confiabilidad de las predicciones realizadas.
- Proponer recomendaciones para optimizar la gestión del stock de especies valoradas y mejorar la planificación de recursos, basadas en los resultados del análisis predictivo y la evaluación de los modelos desarrollados, garantizando así la precisión y confiabilidad de las predicciones realizadas.

2. Capítulo II: Revisión de la Literatura

2.1 Modelos Predictivos en la Administración Pública

La implementación de modelos predictivos en la administración pública ha cobrado relevancia en los últimos años debido a la creciente necesidad de optimizar recursos y mejorar la eficiencia en la prestación de servicios. Estudios como el de López & Ramírez (2018) y Mendoza et al. (2020) han demostrado la efectividad de estos modelos en la gestión de recursos consulares, específicamente en la predicción de la demanda de visas y pasaportes. Sin embargo, todavía existe una brecha en la literatura en cuanto a la predicción de la demanda de especies valoradas en consulados, lo que resalta la necesidad de investigaciones específicas en este ámbito.

Estos modelos permiten anticipar comportamientos futuros a partir de datos históricos, lo cual resulta particularmente útil en entornos donde la demanda de servicios es variable y difícil de predecir. En el contexto consular, la gestión eficiente de recursos como las especies valoradas es esencial para garantizar un servicio oportuno y efectivo.

2.1.1 Aplicación en Consulados y Embajadas

Aunque la literatura específica sobre el uso de modelos predictivos en consulados y embajadas es limitada, existen estudios que destacan su potencial en la gestión de recursos consulares. Un ejemplo es el análisis de la demanda de visas y pasaportes, donde se han aplicado técnicas de minería de datos para predecir la demanda y ajustar los recursos necesarios (López & Ramírez, 2018). Además, la implementación de estos modelos en la gestión de citas y trámites ha mostrado mejoras en la eficiencia y la satisfacción del usuario (Mendoza et al., 2020).

2.1.2 Beneficios de los Modelos Predictivos

Los modelos predictivos ofrecen varios beneficios para la administración pública, incluyendo la mejora en la planificación y gestión de recursos, la capacidad de anticiparse a las necesidades de los ciudadanos y la optimización de procesos operativos (González, 2019). En el contexto consular, esto se traduce en una mejor asignación de especies valoradas, reducción de tiempos de espera y un servicio más eficiente y eficaz para los ciudadanos.

2.2 ¿Qué es Machine Learning “ML”?

El ML es una subdisciplina de la inteligencia artificial que permite a las máquinas aprender y mejorar a partir de la experiencia sin ser programadas explícitamente para cada tarea. Este

aprendizaje se logra a través de algoritmos que analizan datos, identifican patrones y toman decisiones basadas en estos análisis (Mitchell, 1997).

Los algoritmos de ML se clasifican en tres tipos principales:

Aprendizaje supervisado: Utiliza datos etiquetados para entrenar modelos que predicen resultados basados en nuevas entradas.
Aprendizaje no supervisado: Busca patrones y estructuras en datos no etiquetados.
Aprendizaje por refuerzo: Entrena a un agente para tomar decisiones mediante la interacción con un entorno y la recepción de recompensas o castigos.

2.2.1 Clases de Algoritmos de ML

2.2.1.1 Aprendizaje Supervisado

El aprendizaje supervisado es una técnica donde el modelo es entrenado con un conjunto de datos etiquetados, es decir, datos donde el resultado deseado es conocido. Esto permite al modelo aprender a predecir la salida a partir de nuevas entradas basadas en el conocimiento adquirido. Los algoritmos comunes de aprendizaje supervisado incluyen la regresión lineal, la regresión logística, y los árboles de decisión (Hastie et al., 2009).

2.2.1.2 Aprendizaje No Supervisado

El aprendizaje no supervisado se aplica a datos que no tienen etiquetas, y el objetivo es encontrar patrones o estructuras en los datos. Los algoritmos de clustering, como k-means y el análisis de componentes principales (PCA), son ejemplos comunes de técnicas no supervisadas (Hinton & Salakhutdinov, 2006).

2.3 Metodología CRISP-DM para la Creación de Modelos

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un enfoque estándar para abordar proyectos de minería de datos, proporcionando un marco estructurado que guía a los profesionales en la toma de decisiones basadas en datos. Esta metodología se utiliza ampliamente en la industria debido a su enfoque sistemático y adaptable a diferentes tipos de proyectos de análisis de datos y ML (Chapman et al., 2000).

CRISP-DM divide el proceso de minería de datos en seis fases principales, que aseguran que el proyecto se gestione de manera efectiva desde la comprensión del problema hasta la implementación del modelo en un entorno operativo.

2.3.1 Fases de la Metodología CRISP-DM

- **Comprensión del negocio:** En esta fase se define el problema de negocio que se desea resolver y se establecen los objetivos del proyecto.

- **Comprensión de los datos:** Esta fase implica la recopilación y exploración de los datos disponibles. Se busca identificar patrones y relaciones relevantes en los datos que puedan ayudar a abordar el problema de negocio.
- **Preparación de los Datos:** En esta fase, los datos se limpian y transforman para su análisis. Esto puede incluir la eliminación de datos incompletos, la conversión de datos a formatos estándar y la creación de variables derivadas. La preparación de datos es crucial para asegurar la calidad y la precisión de los modelos desarrollados.

- **Modelado:** En esta fase se seleccionan y aplican técnicas de modelado adecuadas. Se entrenan modelos utilizando los datos preparados para encontrar patrones y relaciones.

- **Evaluación:** Esta fase evalúa el rendimiento de los modelos desarrollados para asegurar que cumplen con los objetivos del proyecto. Se pueden utilizar métricas de desempeño y validación cruzada para evaluar la precisión y la utilidad del modelo.

- **Despliegue:** En la fase de despliegue, el modelo se implementa en un entorno operativo para su uso continuo. Esto puede implicar la integración del modelo en sistemas de toma de decisiones y la capacitación del personal para su uso efectivo.

La metodología CRISP-DM es valiosa porque proporciona una guía clara y estructurada para llevar a cabo proyectos de minería de datos de manera eficiente y consistente. Al seguir esta metodología, las organizaciones pueden minimizar errores, mejorar la calidad de sus modelos y tomar decisiones informadas basadas en datos (Shearer, 2000).

Figura 1 Esquema del ciclo CRISP-DM

(https://adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/, 2021)

2.4 Modelos de Predicción

2.4.1 Regresión Lineal

La regresión lineal es una técnica de aprendizaje supervisado que modela la relación entre una variable dependiente continua y una o más variables independientes. Es útil para predecir la demanda de especies valoradas basándose en datos históricos (Montgomery et al., 2012).

2.4.2 Random Forest

Random Forest es un algoritmo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Desarrollado por Leo Breiman y Adele Cutler, este algoritmo combina múltiples árboles de decisión para mejorar la precisión del modelo y evitar el sobreajuste. Cada árbol en el bosque es construido utilizando una muestra aleatoria del conjunto de datos y selecciona una submuestra aleatoria de características para dividir en cada nodo, lo que introduce variabilidad y reduce la correlación entre los árboles individuales.

Funcionamiento del Algoritmo

Bootstrap Aggregation (Bagging): Se crean múltiples subconjuntos de los datos originales mediante muestreo con reemplazo. Cada subconjunto se utiliza para entrenar un árbol de decisión.

Selección de Características: En cada nodo del árbol, se selecciona aleatoriamente un subconjunto de características para determinar la mejor división. Esto reduce la correlación entre los árboles.

Crecimiento de los Árboles: Cada árbol de decisión crece hasta el máximo sin poda, asegurando que cada uno sea diferente.

Promediación: Para problemas de regresión, las predicciones de todos los árboles se promedian; para problemas de clasificación, se toma el voto mayoritario.

Ventajas y Aplicaciones

Robustez: Alta precisión y resistencia al sobreajuste.

Interpretabilidad: Permite evaluar la importancia de cada variable.

Escalabilidad: Eficiente en grandes conjuntos de datos.

2.4.3 Redes Neuronales

Las redes neuronales son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano. Compuestas por capas de nodos (neuronas), estas redes pueden aprender representaciones complejas de datos a través de un proceso iterativo de ajuste de pesos. Las redes neuronales son particularmente efectivas en la detección de patrones no lineales y la captura de relaciones complejas en los datos.

Estructura y Funcionamiento

Capas de la Red: Una red neuronal típica consta de una capa de entrada, una o más capas ocultas, y una capa de salida. Cada neurona en una capa está conectada a todas las neuronas de la siguiente capa.

Propagación hacia Adelante: Los datos de entrada se pasan a través de las capas de la red. En cada neurona, se calculan valores ponderados y se aplican funciones de activación no lineales para producir la salida.

Función de Pérdida: Se calcula el error entre la predicción de la red y el valor real.

Retropropagación: El error se propaga hacia atrás a través de la red, ajustando los pesos mediante el algoritmo de gradiente descendente para minimizar la función de pérdida.

Ventajas y Aplicaciones

Flexibilidad: Capacidad para modelar relaciones no lineales complejas.

Escalabilidad: Adecuado para grandes conjuntos de datos.

Generalización: Buen rendimiento en tareas de predicción y clasificación.

2.4.4 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) es un algoritmo de aprendizaje supervisado utilizado principalmente para problemas de clasificación y regresión. El principio fundamental del K-NN es que una observación se clasifica en función de la mayoría de sus vecinos más cercanos, definidos en términos de distancia en el espacio de características.

Funcionamiento del Algoritmo

Determinación de K: Se selecciona el número de vecinos más cercanos (K). Un valor adecuado de K es crucial para el rendimiento del modelo.

Cálculo de Distancias: Se calcula la distancia (generalmente euclidiana) entre la observación que se va a clasificar y todas las demás observaciones en el conjunto de datos.

Identificación de Vecinos: Se identifican los K vecinos más cercanos.

Asignación de Clase: Para problemas de clasificación, se asigna la clase que es más frecuente entre los K vecinos. Para problemas de regresión, se calcula la media de los valores de los vecinos.

Ventajas y Aplicaciones

Simplicidad: Fácil de implementar y entender.

Eficacia en Datos Locales: Muy efectivo cuando la relación entre las características y la variable de destino es local.

No Paramétrico: No asume una distribución subyacente de los datos.

3. Capítulo III: Marco metodológico

3.1 Metodología

La metodología propuesta se basa en el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), un enfoque estructurado que guiará el desarrollo del proyecto.

3.1.1 Comprensión del negocio

El Consulado del Ecuador en Queens realiza una serie de trámites consulares que requieren el uso de especies valoradas. Estos elementos son fundamentales para el proceso de cobro de tasas en diversos servicios consulares. La gestión eficiente de estas especies es crucial para mantener un servicio consular ágil y eficaz.

A través del análisis de datos históricos, se ha identificado una falta de precisión en las proyecciones de la demanda de especies valoradas, lo que a menudo conduce a desafíos significativos en la gestión del inventario, como el exceso o la escasez de recursos. Esta problemática afecta directamente la eficiencia operativa del consulado, generando potenciales retrasos en la atención y una disminución en la satisfacción de los usuarios. La actual dependencia de estimaciones no precisas para la proyección de demanda donde se observa la necesidad de una metodología más robusta y basada en datos para la planificación y gestión del stock.

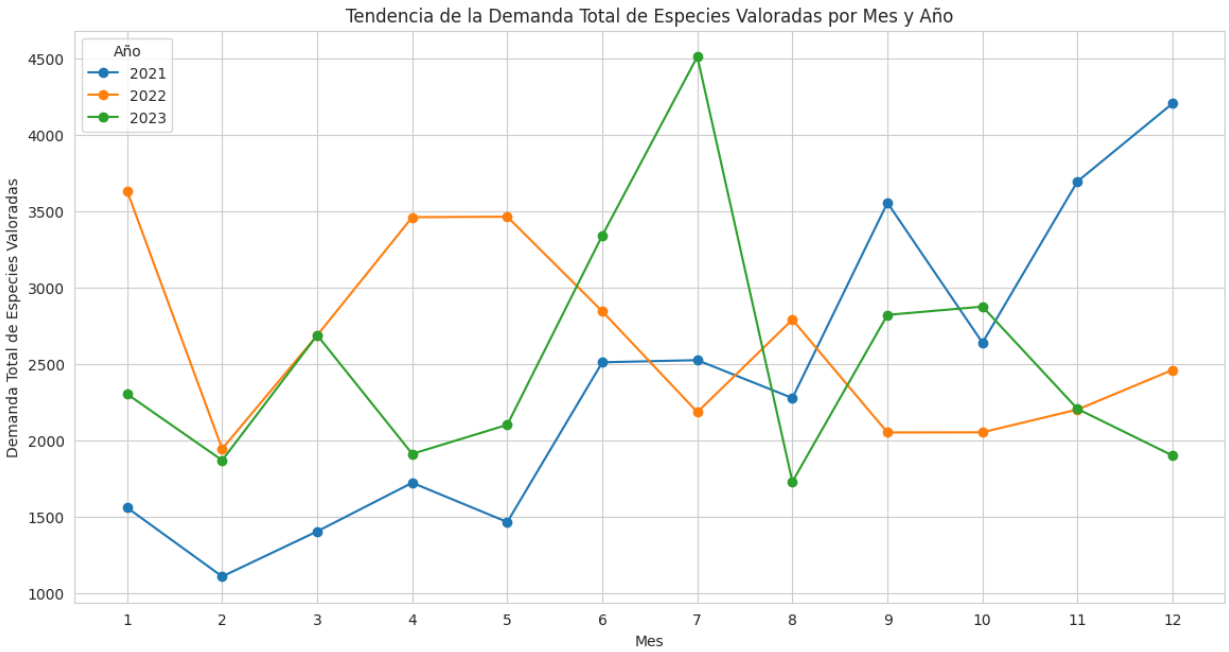

Figura 2 Tendencia de la demanda total de especies valoradas por mes y año.

En la Figura 2, se muestran variaciones significativas en la demanda de especies valoradas a lo largo de varios meses y años. Se observaron picos de demanda en enero de 2022 y julio de 2023, mientras que otros meses mostraron variabilidad en la demanda entre los años estudiados.

Con los resultados de la información observada, se recomienda adoptar un enfoque analítico para la gestión de especies valoradas. Esto no solo mejorará la operación interna del consulado, sino que también aumentará la transparencia y la rendición de cuentas en el uso de recursos. Esto es de suma importancia para la administración pública y fortalece la confianza de la comunidad en los servicios consulares. La implementación de soluciones basadas en datos asegura una asignación más eficiente de recursos y una mejor respuesta a las necesidades de los ciudadanos, contribuyendo a una experiencia más positiva para los usuarios y un mejor cumplimiento de los objetivos institucionales.

3.1.2 Comprensión de los datos

El conjunto de datos proviene de una base de datos estructurada centralizada y correspondiente al Consulado del Ecuador en Queens. Este archivo está compuesto por 18 variables y 90,772 registros. Estas variables contienen información detallada sobre los trámites consulares realizados, las especies valoradas utilizadas, y otras variables relevantes para comprender la gestión de estos recursos. Adicionalmente, es importante destacar que la información otorgada en un archivo Excel incluye únicamente los campos que contienen datos relacionados con el proceso de asignación de especies valoradas.

3.1.2.1 Matriz de Operacionalización de Variables

A continuación, se presenta la Matriz de Operacionalización de Variables, que define las variables con las cuales se realizará el estudio, su tipo, su definición operativa y cómo se medirán.

Tabla 1 Matriz de Operacionalización de Variables

Variable	Tipo	Definición	Medición
IdCentroAdministrativo	Independiente	Identificador del centro administrativo.	Número entero
IdTramite	Independiente	Identificador único del trámite realizado.	Número entero
FechaTramite	Independiente	Fecha en que se registró el trámite.	Fecha en formato YYYY-MM-DD

ValorTramite	Dependiente	Valor monetario asociado al trámite.	Valor numérico en dólares
NumeroTramite	Independiente	Número del trámite realizado.	Número entero
FechaCobro	Independiente	Fecha en que se efectuó el pago por el trámite.	Fecha en formato YYYY-MM-DD
TipoActoConsular	Independiente	Tipo del acto consular realizado.	Categorías predefinidas (ej. Tipo Poderes, Actos Administrativos, Apostillas)
ActoConsular	Independiente	Nombre específico del trámite realizado.	Categorías predefinidas (ej. Poder especial y primera copia, Tarjeta de identificación consular)
FechaAsignacionEspecie	Independiente	Fecha en que se asignó la especie valorada.	Fecha en formato YYYY-MM-DD
Serie	Independiente	Serie de la especie valorada asignada.	Valor alfanumérico
Numero	Independiente	Número de la especie valorada asignada.	Número entero
EspecieValor	Independiente	Valor monetario de la especie fiscal.	Valor numérico en dólares
EspecieNombre	Independiente	Nombre de la especie valorada.	Categorías predefinidas (ej. Timbre $0.50, Timbre $1, Timbre $2, Timbre $5)
NombreTipoActoConsular	Independiente	Nombre del tipo de acto consular realizado.	Categorías predefinidas
MesTramite	Independiente	Mes en que se registró el trámite.	Número entero (1-12)
DiaSemanaTramite	Independiente	Día de la semana en que se registró el trámite.	Número entero (0-6, donde 0 = domingo)

AñoTramite	Independiente	Año en que se registró el trámite.	Número entero
MesAñoTramite	Independiente	Mes y año en que se registró el trámite.	Periodo (YYYY-MM)

3.1.2.2 Exploración de los Datos

Se llevó a cabo una exploración inicial de los datos para comprender su estructura, distribución y características principales.

Visualización de Datos: Se generaron gráficos y tablas para visualizar la distribución de las variables clave, como la cantidad de trámites por mes y año, y la demanda de diferentes especies valoradas.

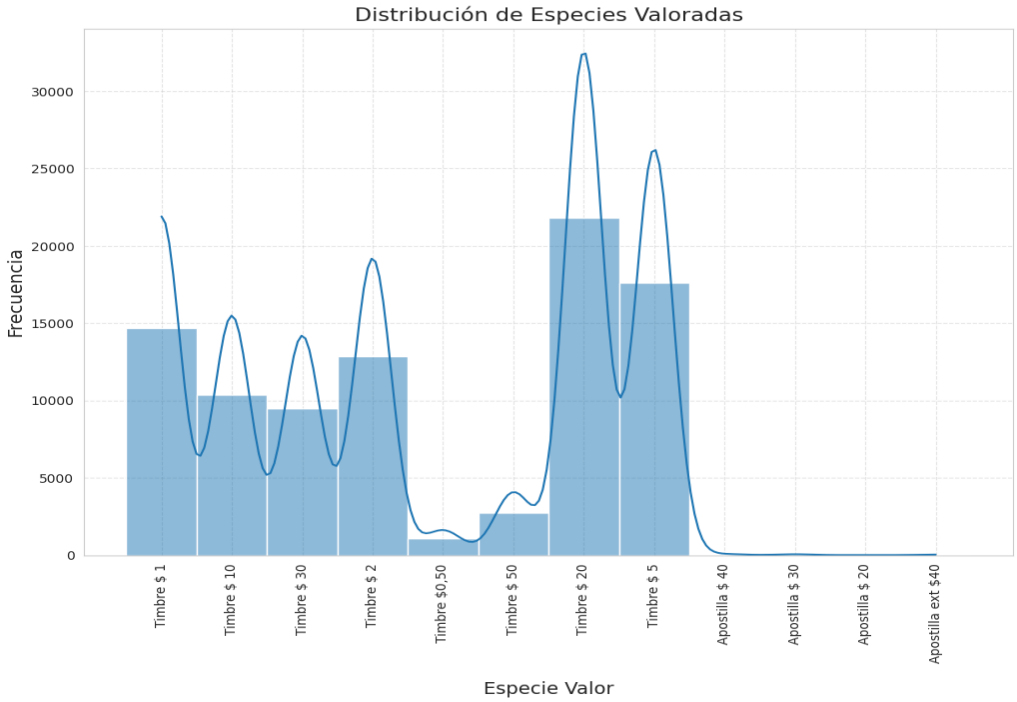

Figura 3 Distribución de especies valoradas

De acuerdo con la Figura 3 existe una alta frecuencia de uso de los timbres de $ 20 y $ 5, esto indica que los valores son críticos para la mayoría de los trámites consulares.

La demanda está distribuida entre diferentes valores de timbres, reflejando la diversidad de trámites consulares que varían en valor. Los timbres de $ 1, $ 2, $ 10, y $ 30 también muestran una notable utilización.

Las apostillas tienen una demanda significativamente menor en comparación con los timbres, lo que podría indicar que los trámites que requieren apostillas son menos frecuentes o están restringidos a ciertos tipos de servicios consulares.

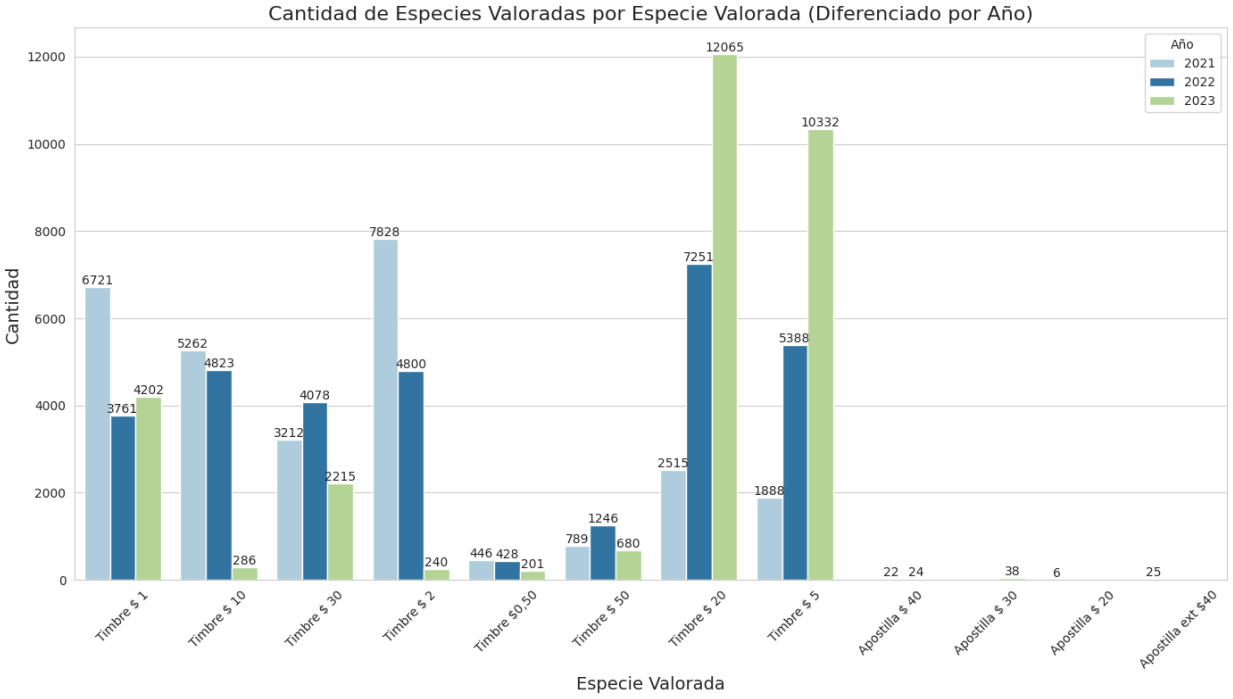

Figura 4 Cantidad de especies valoradas por año.

Como se puede observar en la Figura 4, además de analizar la cantidad total de especies valoradas, se puede ver cómo ha variado el uso de diferentes denominaciones en los años 2021, 2022 y 2023. El análisis muestra que las apostillas tienen una demanda variada.

- **Apostilla $20:** Solo se usó en 2021 (6 veces), indicando una demanda muy baja y puntual.
- **Apostilla $30:** Se usó únicamente en 2023 (38 veces), lo que sugiere una introducción reciente o un uso limitado.
- **Apostilla $40:** Se utilizó en 2022 (22 veces) y en 2023 (24 veces), mostrando una ligera demanda.
- **Apostilla ext $40:** Fue usada en 2021 (25 veces), sin datos para otros años, lo que indica un uso muy específico o discontinuado.

Los timbres muestran patrones distintos:

- **Timbre $0.50:** Tuvo un uso decreciente de 2021 (446) a 2023 (201), indicando una menor necesidad de trámites que requieren este timbre.
- **Timbre $1:** Tuvo un uso fluctuante, disminuyendo en 2022 (3761) respecto a 2021 (6721) pero incrementándose en 2023 (4202).

- **Timbre $2:** Mostró una alta demanda en 2021 (7828) y 2022 (4800), con una drástica reducción en 2023 (240), posiblemente debido a cambios en los tipos de trámites o políticas.
- **Timbre $5:** Tuvo un incremento significativo desde 2021 (1888) a 2023 (10332), sugiriendo un aumento en trámites de mayor valor.
- **Timbre $10:** Mostró una alta demanda en 2021 (5262) y 2022 (4823), pero una reducción drástica en 2023 (286).
- **Timbre $20:** Tuvo un incremento notable desde 2021 (2515) a 2023 (12065), indicando una creciente necesidad de trámites que requieren este valor.
- **Timbre $30:** Presentó una demanda consistente, pero con una disminución en 2023 (2215) respecto a 2021 (3212) y 2022 (4078).
- **Timbre $50:** Tuvo un uso relativamente bajo, con un pico en 2022 (1246) y una caída en 2023 (680).

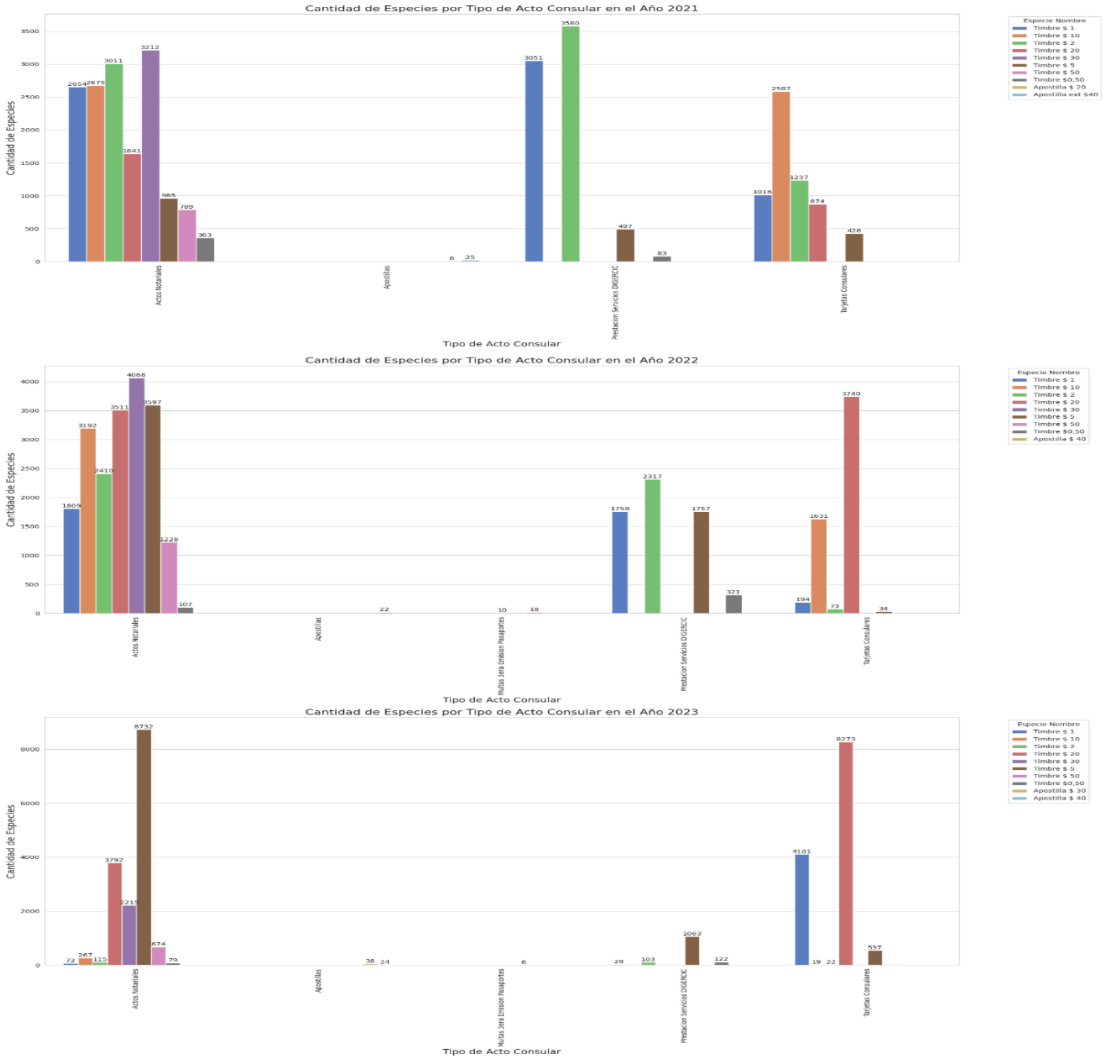

Figura 5 Cantidad de especies valoradas por tipo de trámite.

En los gráficos de la Figura 5, se pueden observar claramente estas tendencias y patrones de uso de las especies valoradas por tipo de acto consular y año. La representación visual ayuda a identificar rápidamente cuáles especies valoradas fueron más demandadas en cada categoría y cómo ha evolucionado su uso a lo largo del tiempo.

Los incrementos significativos en ciertas denominaciones de timbres, como el Timbre $5 en actos notariales en 2023, destacan la necesidad de ajustar el inventario para satisfacer la demanda creciente. De manera similar, la estabilización en el uso de apostillas refleja una demanda constante que puede ser atendida de manera eficiente con una planificación adecuada del stock.

Detección de Anomalías: Se identificaron valores atípicos y anomalías en los datos, que podrían afectar la calidad del análisis y los modelos predictivos.

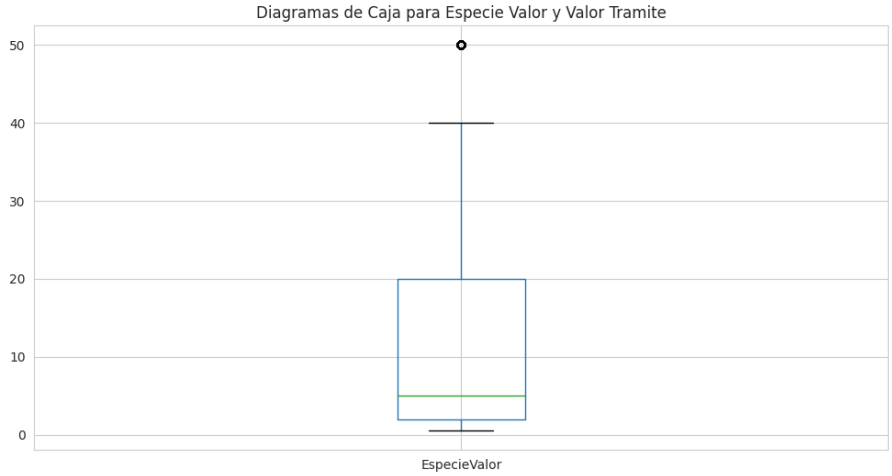

Figura 6 Diagrama de caja - EspecieValor

De acuerdo a la Figura 6, en el análisis de la variable Especie Valor revela varios aspectos importantes sobre la distribución y variabilidad de los valores de las especies utilizadas en los trámites consulares. Con un total de 90,772 registros, se observa que el valor promedio de las especies es de $12.06. Esta media sugiere que, en promedio, las especies valoradas tienen un valor moderado. Sin embargo, la desviación estándar de $11.77 indica una dispersión significativa en los valores, lo que sugiere que hay una amplia variabilidad en los valores de las especies valoradas.

El valor mínimo registrado es de $0.50, representando especies de muy bajo costo, mientras que el valor máximo es de $50.00, correspondiente a las especies de mayor costo. El primer cuartil (25%) muestra que el 25% de las especies tienen un valor de $2.00 o menos, lo que indica una proporción significativa de especies de bajo valor. La mediana de $5.00 sugiere que la mitad de

las especies tienen un valor de $5.00 o menos, mientras que el tercer cuartil (75%) muestra que el 75% de las especies tienen un valor de $20.00 o menos. Solo una cuarta parte de las especies tienen un valor superior a $20.00.

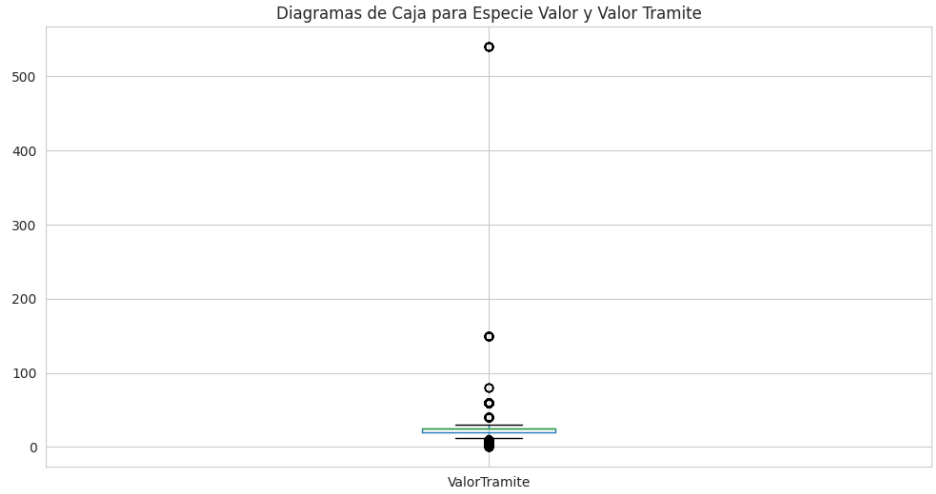

Figura 7 Diagrama de caja – ValorTramite

En la Figura 7, en el análisis de la variable Valor Trámite también revela información importante sobre los costos de los trámites consulares. Con un total de 90,772 registros, el valor promedio de los trámites es de $25.58, sugiriendo un costo moderado en promedio. La desviación estándar de $17.86 indica una amplia variabilidad en los costos de los trámites, con valores que varían considerablemente.

El valor mínimo registrado es de $0.00, indicando la existencia de trámites sin costo, mientras que el valor máximo es de $540.00, representando los trámites de mayor costo. El primer cuartil (25%) muestra que el 25% de los trámites tienen un valor de $20.00 o menos, mientras que la mediana de $25.00 sugiere que la mitad de los trámites tienen un costo de $25.00 o menos. El tercer cuartil (75%) muestra que el 75% de los trámites tienen un valor de $25.00 o menos, lo que indica que solo una cuarta parte de los trámites tienen un costo superior a $25.00.

3.1.2.3 Verificar la Calidad de los Datos

La verificación de la calidad de los datos incluyó varios pasos para asegurar que los datos fueran adecuados para el análisis.

- **Detección de Valores Faltantes**

  Se identificaron y contaron los valores faltantes en el set de datos.

Tabla 2 Valores faltantes y su porcentaje

COLUMNA	VALORES FALTANTES	PORCENTAJE (%)
SERIE	2830	3.12

De acuerdo a la Tabla 2 se puede observar que, en el conjunto de datos se identificó que la columna "Serie" presentaba un 3.12% de valores faltantes.

- **Detección de Duplicados**

  El análisis reveló que no existen registros duplicados en el conjunto de datos. Esto indica que cada registro en el DataFrame es único, y no hay entradas repetidas que podrían haber comprometido la integridad del análisis.

- **Consistencia de los Datos**

  Para garantizar la consistencia de los datos en el conjunto de datos, se llevaron a cabo varias acciones, incluyendo la conversión de fechas y la corrección de formatos y rangos. Estas acciones son esenciales para asegurar que los datos sean coherentes y adecuados para el análisis y la modelización.

  La correcta interpretación y manipulación de datos temporales es crucial en cualquier análisis de datos. Por lo tanto, se realizó la conversión de las columnas de fechas al formato datetime de pandas, lo que permite un manejo más efectivo y preciso de la información temporal.

- **Creación de columnas adicionales para el análisis:**
  En el análisis de datos, la creación de nuevas columnas derivadas de las existentes puede proporcionar información adicional y facilitar un análisis más detallado y significativo. Estas columnas adicionales permiten explorar patrones temporales y otros aspectos específicos de los datos que pueden no ser evidentes en las columnas originales.

  Para enriquecer el conjunto de datos del Consulado del Ecuador en Queens y facilitar un análisis más exhaustivo, se crearon columnas adicionales a partir de las fechas de los

trámites consulares. Las nuevas columnas incluyen el mes y el día de la semana en que se realizaron los trámites.

Estos pasos aseguraron que los datos utilizados para el análisis fueran de alta calidad y estuvieran libres de errores significativos que pudieran sesgar los resultados. Se identificaron y manejaron valores faltantes, se eliminaron duplicados y se corrigieron inconsistencias en fechas y otros datos críticos.

3.1.2.4 Preparación de los datos

3.1.2.4.1 Seleccionar los Datos

La selección de las variables más relevantes del conjunto de datos es un paso crucial para el análisis y modelamiento, asegurando que solo la información pertinente se utilice para construir modelos predictivos efectivos. El proceso de selección de datos se basó en un Análisis Exploratorio de Datos (EDA) detallado, que incluyó visualizaciones y estadísticas descriptivas para identificar las variables clave.

Basado en el análisis exploratorio de datos, se seleccionaron las siguientes variables clave para el análisis y modelamiento:

Tabla 3Variables seleccionadas para el análisis y modelado

Tipo de la Variable	Variables Originales	Variables Usadas
Numéricas	ValorTramite	ValorTramite
	EspecieValor	EspecieValor
	MesTramite	MesAñoTramite (derivada de FechaTramite)
	AñoTramite	AñoTramite (derivada de FechaTramite)
	NumeroTramite	
	IdTramite	

	Numero	
	IdCentroAdministrativo	
	DiaSemanaTramite	
Categóricas	EspecieNombre	EspecieNombre
	TipoActoConsular	NombreTipoActoConsular (derivada de TipoActoConsular)
	ActoConsular	
	Serie	
Fechas	FechaTramite	FechaTramite (convertida a datetime)
	FechaCobro	
	FechaAsignacionEspecie	

De acuerdo a la Tabla 3, se seleccionan estas variables porque proporcionan la información necesaria para construir modelos predictivos que pueden anticipar la demanda futura de especies valoradas basándose en patrones históricos y otras características relevantes.

3.1.2.4.2 Limpiar los Datos

La limpieza de datos fue un paso crucial para asegurar la calidad y precisión del análisis. Este proceso incluyó varias actividades fundamentales para preparar los datos para el análisis y modelamiento:

- **Imputación de Valores Faltantes:**
  - **Identificación e imputación de Valores Faltantes:**

Se identificaron y cuantificaron los valores faltantes en el conjunto de datos. La columna "Serie" presentó un 3.12% de valores faltantes. Se determinó que estos valores correspondían a especies valoradas que no tenían una serie asignada.

Para mantener la integridad del conjunto de datos y evitar la pérdida de información, se realizó la imputación de estos valores faltantes con el valor "SinSerie". Esto permitió conservar completamente el conjunto de datos sin introducir sesgos significativos.

- **Detección y Eliminación de Duplicados:**
  - **Identificación de Registros Duplicados:**

    Al realizar el análisis del conjunto de datos para detectar la presencia de registros duplicados, se llega a la conclusión de que no existían registros duplicados en el conjunto de datos, lo que indica que cada registro en el DataFrame es único.

  - **Acción Tomada:**

    Dado que no se encontraron duplicados, no fue necesario realizar ninguna eliminación. Esto confirmó la calidad y unicidad de los datos recolectados.

- **Detección y Tratamiento de Outliers**

  El tratamiento de outliers es esencial para asegurar la calidad y precisión del análisis. Se realizó un análisis detallado de los datos con y sin outliers:

  Resultados con Outliers:

	EspecieValor	ValorTramite
count	90772.000000	90772.000000
mean	12.055783	25.578951
std	11.770272	17.857635
min	0.500000	0.000000
25%	2.000000	20.000000
50%	5.000000	25.000000
75%	20.000000	25.000000
max	50.000000	540.000000

Figura 8 Outliers en las variables EspecieValor y ValorTramite

De acuerdo a la Figura 8, el conjunto de datos con outliers muestra una alta variabilidad, especialmente en la variable ValorTramite, cuyo valor máximo es 540. Esto sugiere la presencia de valores extremos que pueden influir significativamente en el análisis.

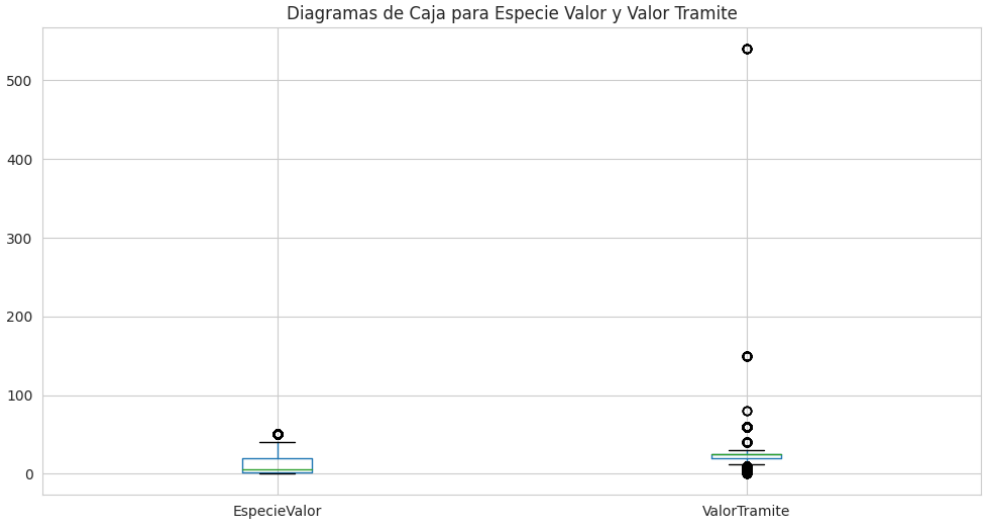

Figura 9 Diagrama de cajas con outliers para EspecieValor y ValorTramite

Resultados sin Outliers:

	EspecieValor	ValorTramite
count	90726.000000	90726.000000
mean	12.039410	25.464283
std	11.748737	16.680536
min	0.500000	0.000000
25%	2.000000	20.000000
50%	5.000000	25.000000
75%	20.000000	25.000000
max	50.000000	80.000000

Figura 10 Sin outliers en las variables EspecieValor y ValorTramite

Tras la eliminación de outliers, se puede observar en la Figura 10 una reducción en la variabilidad de los datos, lo que resulta en una desviación estándar menor para ValorTramite. El valor máximo de ValorTramite se reduce a 80, mostrando que los valores extremos han sido eliminados.

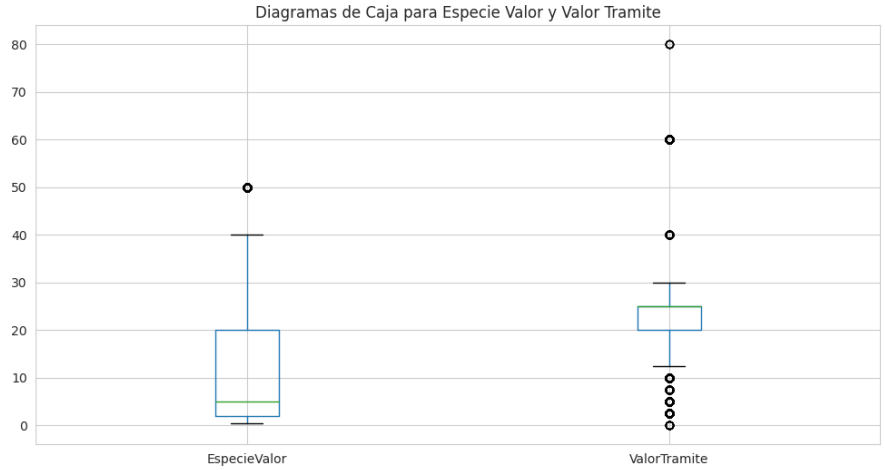

Figura 11 Diagrama de cajas sin outliers para EspecieValor y ValorTramite

La eliminación de outliers mejora la consistencia y precisión del análisis al reducir la variabilidad y eliminar influencias extremas sin comprometer la representatividad del conjunto de datos. Estos pasos aseguraron que los datos utilizados para el análisis fueran de alta calidad y estuvieran libres de errores significativos que pudieran sesgar los resultados.

- **Generación de Nuevas Columnas:**
  Para enriquecer el análisis, se crearon columnas adicionales derivadas de las fechas de los trámites consulares.

  - **Mes y Año del Trámite:** Se extrajeron y crearon columnas para el mes (MesTramite) y el año (AñoTramite) de los trámites.

  - **Día de la Semana del Trámite:** Se generó una columna para el día de la semana (DiaSemanaTramite) en que se realizaron los trámites.

3.1.2.4.3 Estructurar los Datos

La estructuración de los datos se enfocó en organizar los datos en un formato adecuado para el análisis y modelamiento. Este paso es crucial para asegurar que los datos estén listos para ser procesados por modelos estadísticos y de ML, y que las relaciones temporales y categóricas se manejen correctamente

- **Conversión de Formatos:**

La conversión de fechas a un formato datetime adecuado fue uno de los primeros y más importantes pasos en la estructuración de los datos. Las columnas de fechas (FechaTramite, FechaCobro, FechaAsignacionEspecie) fueron convertidas al formato datetime de pandas. Esta conversión es esencial para cualquier análisis temporal preciso, ya que permite realizar operaciones y manipulaciones de tiempo, como calcular diferencias, extraer componentes temporales (año, mes, día de la semana) y agrupar datos por intervalos de tiempo.

La correcta conversión de fechas no solo facilita el manejo de los datos temporales, sino que también permite utilizar métodos avanzados de pandas para la manipulación de fechas y tiempos, mejorando la precisión y la eficacia del análisis temporal. Este proceso asegura que todas las fechas estén en un formato estándar, evitando errores que pueden surgir de la diversidad de formatos de fecha.

- **Codificación de Variables Categóricas:**
  Para facilitar el análisis y modelado, se realizó la codificación de variables categóricas. Las variables categóricas pueden ser difíciles de manejar directamente por muchos algoritmos de ML, por lo que se utilizan técnicas como One-Hot Encoding y Label Encoding para convertir estas variables en formatos numéricos.

  One-Hot Encoding se utilizó para variables categóricas que no tienen un orden intrínseco y donde cada categoría es igualmente importante. Esta técnica convierte cada categoría en una columna binaria (0 o 1), indicando la presencia o ausencia de la categoría en cada fila. Esto es especialmente útil para algoritmos que no pueden manejar variables categóricas directamente, como la regresión lineal y las redes neuronales, ya que evita la creación de una jerarquía falsa entre categorías.

  Label Encoding se aplicó a variables categóricas donde las categorías tienen un orden intrínseco o donde la cantidad de categorías es muy alta. Esta técnica asigna un número entero único a cada categoría, simplificando el manejo de estas variables en algunos algoritmos de ML que pueden interpretar los valores como ordinales. La reducción de dimensionalidad en comparación con One-Hot Encoding es otra ventaja significativa de esta técnica.

- **Conversión de Variables Temporales a Componentes**

  Además de convertir las fechas al formato datetime, se extrajeron componentes adicionales como el mes (MesTramite), el año (AñoTramite), y el día de la semana (DiaSemanaTramite).

  Estas nuevas columnas permiten realizar análisis estacionales y temporales más detallados, facilitando la identificación de patrones y tendencias a lo largo del tiempo. La extracción de estos componentes temporales es crucial para entender cómo varían los datos en diferentes periodos y para hacer predicciones más precisas.

3.1.2.4.4 Exploración Previo al Modelamiento de los Datos

Antes de proceder con el modelamiento, se realizó una exploración adicional de los datos preparados para asegurar su adecuación. Esta etapa es crucial para entender mejor las características y relaciones dentro del conjunto de datos y para asegurar que los modelos predictivos se construyan sobre una base sólida y bien comprendida

- **Matriz de Correlación:**

  Se calculó y visualizó una matriz de correlación para identificar relaciones significativas entre las variables. El mapa de calor de la matriz de correlación proporciona una representación visual de estas relaciones, facilitando la identificación de variables que están fuertemente correlacionadas.

  Esto es fundamental para la construcción de modelos predictivos robustos, ya que permite seleccionar las variables más relevantes y evitar problemas de multicolinealidad.

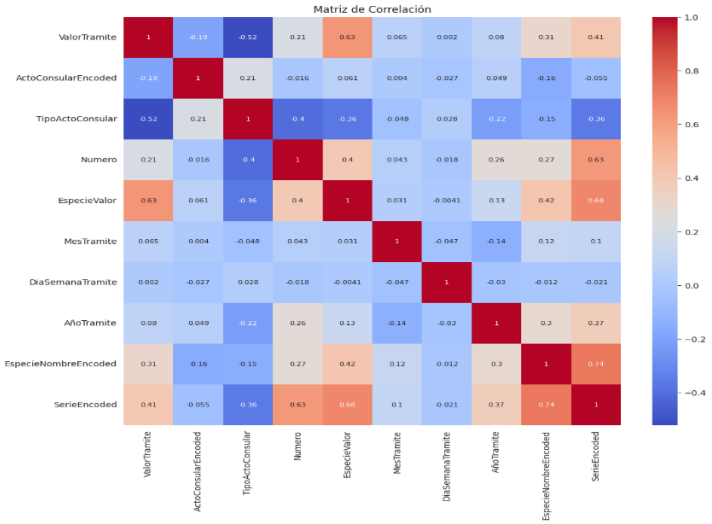

Figura 12 Matriz de correlación

De acuerdo con los resultados de la matriz de correlación observados en la Figura 12, algunas observaciones clave son:

- **ValorTramite y EspecieValor**

  La correlación de 0.658940 entre el valor del trámite y el valor de la especie indica una fuerte relación positiva. Esto significa que a medida que aumenta el valor de la especie valorada, también lo hace el valor del trámite asociado.

  Esta relación es lógica e intuitiva, ya que las especies de mayor costo generalmente se asocian con trámites de mayor valor monetario. Esta información puede ser útil para prever la demanda de especies más costosas y planificar los recursos en consecuencia.

- **ValorTramite y TipoActoConsular**

  La correlación de -0.565424 entre el valor del trámite y el tipo de acto consular muestra una relación negativa significativa. Esto sugiere que ciertos tipos de actos consulares están asociados con trámites de menor valor.

  Es esencial identificar qué tipos específicos de actos consulares tienden a tener valores de trámite más bajos, ya que esto puede influir directamente en la planificación de inventarios y recursos. Con esta información, se pueden desarrollar estrategias para optimizar el uso de especies valoradas en función del tipo de acto consular.

- **ValorTramite y SerieEncoded**

  La correlación de 0.432406 entre el valor del trámite y la serie codificada indica una relación positiva. Esto sugiere que ciertas series están asociadas con trámites de mayor valor.

  Esta correlación podría reflejar políticas o prácticas específicas del consulado en cuanto a la asignación de series para diferentes tipos de trámites. Comprender estas prácticas puede ayudar a mejorar la gestión del inventario y asegurar que las series de mayor valor estén disponibles para los trámites que lo requieran.

- **EspecieValor y SerieEncoded**

Una alta correlación positiva de 0.680926 entre el valor de la especie y la serie codificada sugiere que ciertas series están consistentemente asociadas con las especies valoradas.

- **EspecieNombreEncoded y SerieEncoded**
  La fuerte correlación de 0.736747 entre el nombre de la especie codificada y la serie codificada indica una relación sistemática en cómo se asignan los nombres de las especies a las series. Esto puede ser relevante para analizar y predecir patrones de demanda basados en la serie de la especie, mejorando la precisión de las proyecciones y la planificación de recursos.

- **ValorTramite y AñoTramite**
  Aunque la correlación positiva de 0.085428 entre el valor del trámite y el año del trámite es baja, sugiere una ligera tendencia al aumento del valor de los trámites con el tiempo. Esto podría reflejar cambios en las tarifas o en la naturaleza de los trámites consulares a lo largo de los años. Esta tendencia es importante para la planificación a largo plazo y la adaptación a posibles aumentos en los costos de los trámites.

- **ValorTramite y MesTramite**
  La correlación de 0.072304 entre el valor del trámite y el mes del trámite indica una correlación positiva baja. Esto sugiere que puede haber variaciones estacionales en el valor de los trámites. Este hallazgo puede ser útil para la planificación estacional y la gestión de recursos, asegurando que el consulado esté preparado para los cambios en la demanda a lo largo del año.

- **ActoConsularEncoded y TipoActoConsular**
  La correlación positiva de 0.178827 entre el acto consular codificado y el tipo de acto consular indica que ciertos actos consulares están relacionados con tipos específicos de actos consulares. Este conocimiento puede ayudar a mejorar la categorización y el análisis de los trámites, facilitando una gestión más eficiente y precisa de los recursos consulares.

Estos insights proporcionan una base sólida para seleccionar y utilizar las variables más relevantes en los modelos predictivos, asegurando que las relaciones significativas sean capturadas efectivamente.

4. Capítulo IV: Resultados y Análisis

En esta sección, se presentan los resultados obtenidos del análisis de los datos históricos de las especies valoradas utilizadas por el Consulado del Ecuador en Queens durante los años 2021, 2022 y 2023. Además, se evaluará la efectividad de los modelos predictivos desarrollados y se propondrán recomendaciones específicas para la optimización de la gestión del stock.

4.1 Datos Recuperados de las Especies Valoradas

A continuación, se presentan tablas detalladas que resumen los datos de las especies valoradas utilizadas, así como los saldos iniciales y finales para cada año. Estas tablas ofrecen una visión clara de cómo se gestionaron las especies valoradas en el Consulado del Ecuador en Queens durante los años 2021, 2022 y 2023, permitiendo identificar patrones de uso y posibles áreas de mejora en la planificación de recursos.

Tabla 4 Resumen de especies valoradas del año 2021.

ESPECIE VALORADA	2021			
	Saldo Inicial	Enviadas por Dirección Administrativa	Especies valoradas Utilizadas	Saldo Final
Timbre $ 0,50	787	0	446	341
Timbre $ 1	3554	5000	6721	1833
Timbre $ 2	8501	0	7828	673
Timbre $ 5	25	2000	1888	137
Timbre $ 10	3571	3800	5262	2109
Timbre $ 20	780	2350	2515	615
Timbre $ 30	4764	0	3212	1552
Timbre $ 50	3227	0	789	2438
Apostilla $ 20, $ 40	20	35	31	24

Tabla 5 Especies valoradas enviadas por la Dirección Administrativa en 2021

Mes	Memorando	Timbre $ 0,50	Timbre $ 1	Timbre $ 2	Timbre $ 5	Timbre $ 10	Timbre $ 20	Timbre $ 30	Timbre $ 50	Apostilla $ 20, $ 40
Julio	MREMH-AGQUEENS-2021-0094-M	0	1000	0	2000	1000	2000	0	0	35
Octubre	MREMH-AGQUEENS-2021-0520-M	0	4000	0	0	2800	350	0	0	0

De acuerdo a la Tabla 5, se puede observar lo siguiente:

- En 2021, se realizaron dos envíos significativos de especies valoradas en julio y octubre con los memorandos MREMH-AGQUEENS-2021-0094-M y MREMH-AGQUEENS-2021-0520-M, respectivamente.
- Los envíos incluyeron Timbres $1, $5, $10, y $20, con un notable incremento en las cantidades enviadas en octubre, especialmente para Timbres $1 y $10.
- El uso elevado de Timbres $1 y $10 indica una alta demanda en esos meses, reflejando posiblemente un aumento en los trámites consulares correspondientes.

Tabla 6 Resumen de especies valoradas del año 2022.

ESPECIE VALORADA	2022			
	Saldo Inicial	Enviadas por Dirección Administrativa	Especies valoradas Utilizadas	Saldo Final
Timbre $ 0,50	341	350	428	263
Timbre $ 1	1833	7000	3761	5072
Timbre $ 2	673	4800	4800	673
Timbre $ 5	137	13250	5388	7999
Timbre $ 10	2109	3000	4823	286
Timbre $ 20	615	7200	7251	564
Timbre $ 30	1552	3000	4078	474
Timbre $ 50	2438	0	1246	1192
Apostilla $ 20, $ 40	24	30	22	32

Tabla 7 Especies valoradas enviadas por la Dirección Administrativa en 2022

Mes	Memorando	Timbre $ 0,50	Timbre $ 1	Timbre $ 2	Timbre $ 5	Timbre $ 10	Timbre $ 20	Timbre $ 30	Timbre $ 50	Apostilla $ 20, $ 40
Febrero	MREMH-AGQUEENS-2022-0024-M	0	0	800	1200	2000	1400	1000	0	0
Marzo	MREMH-AGQUEENS-2022-0164-M	200	2000	4000	250	0	0	0	0	0
Abril	MREMH-AGQUEENS-2022-0406-M Y MREMH-AGQUEEN	0	0	0	1000	1000	2000	600	0	0

	S-2022-0383-M									
Julio	MREMH-AGQUEENS-2022-0479-M	0	5000	0	700	0	300	200	0	0
Septiembre	MREMH-AGQUEENS-2022-0719-M	150	0	0	6100	0	1000	1200	0	0
Octubre	MREMH-AGQUEENS-2022-0853-M	0	0	0	4000	0	2500	0	0	30

De acuerdo a la Tabla 7, se puede observar lo siguiente:

- En 2022, la Dirección Administrativa realizó envíos en febrero, abril, mayo, julio, septiembre y octubre con varios memorandos, incluyendo MREMH-AGQUEENS-2022-0024-M, MREMH-AGQUEENS-2022-0164-M, MREMH-AGQUEENS-2022-0406-M, y otros.
- Estos envíos fueron más frecuentes y de mayor volumen en comparación con 2021, destacando un aumento en la proactividad para mantener el stock adecuado.
- Los envíos incluyeron una amplia variedad de denominaciones, con grandes cantidades de Timbres $5 y $20 enviados a mediados de año, sugiriendo una mayor demanda durante esos meses.

Tabla 8 Resumen de especies valoradas del año 2023.

ESPECIE VALORADA	**2023**			
	Saldo Inicial	Enviadas por Dirección Administrativa	Especies valoradas Utilizadas	Saldo Final
Timbre $ 0,50	263	70	201	132
Timbre $ 1	5072	0	4202	870
Timbre $ 2	673	0	240	433
Timbre $ 5	7999	4000	10332	1667
Timbre $ 10	286	0	286	0
Timbre $ 20	564	17800	12065	6299
Timbre $ 30	474	4000	2215	2259
Timbre $ 50	1192	0	680	512
Apostilla $ 20, $ 40	32	50	62	20

Tabla 9 Especies valoradas enviadas por la Dirección Administrativa en 2023

Mes	Memorando	Timbre $ 0,50	Timbre $ 1	Timbre $ 2	Timbre $ 5	Timbre $ 10	Timbre $ 20	Timbre $ 30	Timbre $ 50	Apostilla $ 20, $ 40
Marzo	MREMH-AGQUEENS-2023-0284-M	70	0	0	0	0	0	0	0	50
Agosto	MREMH-AGQUEENS-2023-0682-M Y MREMH-AGQUEENS-2023-0674-M	0	0	0	4000	0	4000	2000	0	0
Noviembre	MREMH-AGQUEENS-2023-01042-M	0	0	0	0	0	5800	0	0	0

De acuerdo a la Tabla 9, se puede observar lo siguiente:

- En 2023, la Dirección Administrativa realizó varios envíos de especies valoradas en los meses de marzo, agosto y noviembre, con los memorandos MREMH-AGQUEENS-2023-0284-M, MREMH-AGQUEENS-2023-0682-M, MREMH-AGQUEENS-2023-0674-M, y MREMH-AGQUEENS-2023-01042-M.
- Los envíos en marzo incluyeron una cantidad moderada de Timbres $0,50 y Apostillas $20, $40.
- En agosto, se observaron envíos masivos de Timbres $5, $20, y $30, lo que refleja una alta demanda durante ese periodo.
- El envío en noviembre fue significativo, con una gran cantidad de Timbres $20, lo que sugiere una previsión para cubrir una demanda elevada hacia el final del año

4.2 Evaluación de la Efectividad de los Modelos Predictivos

Para evaluar la efectividad de los modelos predictivos, se compararon las predicciones de demanda con las especies valoradas realmente utilizadas en cada año. A continuación, se presentan los resultados de esta comparación para cada modelo utilizado.

4.2.1 Regresión lineal

La regresión lineal se utilizó como modelo inicial debido a su simplicidad y capacidad para proporcionar una línea base clara para la comparación con modelos más complejos. Este modelo ayuda a entender las relaciones lineales entre las variables independientes y la variable dependiente (EspecieNombreEncoded), ofreciendo una interpretación clara de los coeficientes de cada variable.

- **Selección de Características**

  Las características seleccionadas para el modelo incluyeron AñoTramite, TrimestreTramite, MesTramite, DiaTramite, ActoConsularEncoded, EspecieValor, ValorTramite y SerieEncoded. Estas variables fueron elegidas debido a su relevancia potencial en la determinación del conteo de especies.

```
# Selección de características agrupadas
features = [
    'AñoTramite','TrimestreTramite','MesTramite', 'DiaTramite',
    'ActoConsularEncoded', 'EspecieValor',
    'ValorTramite',
    'SerieEncoded'
]

X = df_grouped[features]
y = df_grouped['EspecieNombreEncoded']
```

Figura 13 Selección de características - Regresión lineal

- **División de Datos**

  Los datos fueron divididos en conjuntos de entrenamiento y prueba, utilizando un 80% de los datos para entrenamiento y un 20% para prueba. Esta división es crucial para evaluar de manera justa el rendimiento del modelo en datos no vistos durante el entrenamiento.

```
# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figura 14 División de datos - Regresión lineal

- **Entrenamiento**

  El modelo de regresión lineal se entrenó utilizando el conjunto de entrenamiento. La regresión lineal se eligió debido a su simplicidad y capacidad para proporcionar una línea base inicial para las predicciones. Este método es útil para entender las relaciones lineales entre la variable dependiente y las variables independientes (Montgomery, Peck, & Vining, 2012).

```
# Crear el modelo de regresión lineal
modelo = LinearRegression()

# Entrenar el modelo
modelo.fit(X_train, y_train)

# Hacer predicciones
y_train_pred = modelo.predict(X_train)
y_test_pred = modelo.predict(X_test)
```

Figura 15 Entrenamiento - Regresión lineal

- **Evaluación**

  Se evaluó el rendimiento del modelo utilizando el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2). Estos métodos son comunes para evaluar modelos de regresión, proporcionando una medida de la precisión del modelo y su capacidad para explicar la variabilidad en los datos (Draper & Smith, 1998). Los resultados obtenidos indican que el modelo tiene una capacidad moderada para predecir la demanda de especies valoradas.

```
# Evaluar el modelo
mse_train = mean_squared_error(y_train, y_train_pred)
r2_train = r2_score(y_train, y_train_pred)
mse_test = mean_squared_error(y_test, y_test_pred)
r2_test = r2_score(y_test, y_test_pred)

print(f'MSE (train): {mse_train:.2f}')
print(f'R² (train): {r2_train:.2f}')
print(f'MSE (test): {mse_test:.2f}')
print(f'R² (test): {r2_test:.2f}')
```

Figura 16 Evaluación - Regresión lineal

El modelo fue evaluado utilizando métricas como el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2).

Resultados Actualizados:

- **MSE (train):** 2.24
- **R^2 (train):** 0.44
- **MSE (test):** 2.37
- **R^2 (test):** 0.43

Estos resultados indican que el modelo explica aproximadamente el 44% de la variabilidad en los conteos de especies en el conjunto de entrenamiento y el 43% en el conjunto de prueba, lo cual sugiere que el modelo tiene una capacidad predictiva moderada pero consistente.

- **Análisis de Resultados**

Para profundizar en el análisis del modelo, se generaron DataFrames con las predicciones y los valores reales tanto del conjunto de entrenamiento como del conjunto de prueba:

```
# Crear DataFrames con los resultados
resultados_train = pd.DataFrame({'Real': y_train, 'Prediccion': y_train_pred})
resultados_test = pd.DataFrame({'Real': y_test, 'Prediccion': y_test_pred})

# Mostrar algunos resultados del conjunto de entrenamiento
print("\nResultados del conjunto de entrenamiento:")
print(resultados_train.head())

# Mostrar algunos resultados del conjunto de prueba
print("\nResultados del conjunto de prueba:")
print(resultados_test.head())
```

Figura 17 Análisis de resultados - Regresión lineal

- **Resultados del conjunto de entrenamiento:**

Tabla 10 Conjunto de entrenamiento - Regresión lineal

REAL	PREDICCIÓN
5	5.401099
7	8.685748
9	7.199900
10	6.858153
6	5.322486

- **Resultados del conjunto de prueba:**

Tabla 11 Conjunto de pruebas - Regresión lineal

REAL	PREDICCIÓN
6	6.140818
8	7.816792
9	8.070810
7	8.151927
7	7.712614

Estos resultados muestran que el modelo de regresión lineal tiene un rendimiento razonable, aunque sigue habiendo algunos casos en los que las predicciones difieren de los valores reales. No obstante, el incremento en el coeficiente de determinación (R^2) indica que el modelo ha mejorado su capacidad para explicar la variabilidad en los datos. Sin embargo, es posible que explorar modelos más complejos o aplicar técnicas adicionales de ajuste pueda mejorar aún más la precisión del modelo.

4.2.2 Random Forest

Debido a las conclusiones obtenidas con el modelo de regresión lineal, se identificó la necesidad de explorar modelos más complejos y avanzados para mejorar la precisión de las predicciones. Uno de los modelos considerados es el Random Forest, una técnica de aprendizaje automático conocida por su capacidad de manejar grandes conjuntos de datos, capturar relaciones no lineales y reducir el riesgo de sobreajuste. Random Forest crea múltiples árboles de decisión durante el entrenamiento y combina sus resultados para obtener predicciones más precisas y robustas (Breiman, 2001).

- **Selección de Variables y División de Datos**

Al igual que en el modelo de regresión lineal, se seleccionaron las siguientes variables como características predictivas:

Tabla 12 Características - Random Forest

CARACTERÍSTICAS PREDICTIVAS
ACTOCONSULARENCODED
MESTRAMITE
AÑOTRAMITE
ESPECIEVALOR
VALORTRAMITE
DIATRAMITE
TRIMESTRETRAMITE
SERIEENCODED
ACTOCONSULARENCODED

La variable objetivo seleccionada fue:

- EspecieNombre

El conjunto de datos se dividió en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) para evaluar el rendimiento del modelo. Esta división se la realiza después de varias pruebas realizadas por lo que esto asegura que el modelo se entrena en una porción de los datos y se prueba en datos que no ha visto previamente, lo cual es crucial para evaluar su capacidad de generalización.

```
# Selección de características agrupadas
features = [
    'AñoTramite','TrimestreTramite','MesTramite', 'DiaTramite',
    'ActoConsularEncoded', 'EspecieValor', 'ValorTramite', 'SerieEncoded']

X = df_grouped[features]
y = df_grouped['EspecieNombre']

# Codificación de la variable objetivo
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# Estandarización de las características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled,
                                                    y_encoded,
                                                    test_size=0.2,
                                                    random_state=42)
```

Figura 18 Selección de variables y división de datos - Random Forest

- **Entrenamiento del Modelo**

  El modelo de Random Forest se entrenó utilizando el conjunto de entrenamiento. Se configuró con 100 árboles de decisión (n_estimators=100) tras realizar una prueba inicial con diferentes valores de n_estimators para balancear entre precisión y tiempo de entrenamiento.

  La elección de Random Forest se basó en su robustez al manejar grandes volúmenes de datos y su capacidad para evitar el sobreajuste, lo que lo hace adecuado para este tipo de problemas con datos diversos y no lineales.

  Además, se optó por no realizar poda en los árboles para asegurar que cada árbol capturara la mayor cantidad posible de variabilidad en los datos.

```
# Construcción del modelo de Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

Figura 19 Entrenamiento del modelo - Random Forest

De acuerdo a la Figura 19, se configuró el modelo con 100 árboles de decisión (n_estimators=100) para equilibrar entre precisión y tiempo de entrenamiento. Esta configuración permite capturar variaciones y patrones complejos en los datos sin incurrir en un costo computacional excesivo.

- **Evaluación del Modelo**

  El rendimiento del modelo se evaluó utilizando la precisión (accuracy) y el reporte de clasificación (classification report). La precisión mide la proporción de predicciones correctas, mientras que el reporte de clasificación proporciona detalles sobre la precisión, el recall y el F1-score para cada clase, ofreciendo una visión completa del rendimiento del modelo (Kuhn & Johnson, 2013).

```
# Evaluación del modelo
y_pred = rf_model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred, target_names=le.classes_)

print("Accuracy:", accuracy)
print("Classification Report:\n", report)
```

Figura 20 Evaluación del modelo - Random Forest

- **Resultados:**
  - **Accuracy: 0.9968**
  - **Classification Report:** se pueden verificar en la Tabla 13, mostrando que el modelo mantiene un rendimiento excelente en la mayoría de las clases, con una precisión promedio ponderada de casi 1.00.

Tabla 13 Resultados del modelo - Random Forest

CLASE	PRECISION	RECALL	F1-SCORE	SOPORTE
APOSTILLA $ 20	1.00	1.00	1.00	2
APOSTILLA $ 30	1.00	0.80	0.89	5
APOSTILLA $ 40	1.00	1.00	1.00	7
TIMBRE $ 1	1.00	1.00	1.00	208

TIMBRE $ 10	0.99	1.00	1.00	333
TIMBRE $ 2	0.99	0.99	0.99	184
TIMBRE $ 20	1.00	1.00	1.00	295
TIMBRE $ 30	1.00	1.00	1.00	339
TIMBRE $ 5	1.00	0.99	0.99	266
TIMBRE $ 50	1.00	1.00	1.00	154
TIMBRE $0,50	1.00	0.97	0.99	70

- **Alta Precisión y Recall**

  Aunque la mayoría de las clases muestran una precisión y un recall cercanos al 100%, algunas clases, como "Apostilla $ 30," muestran una ligera disminución en el recall (80%), lo que indica que el modelo sigue siendo altamente efectivo en la mayoría de las clases, pero podría tener dificultades para identificar correctamente todas las instancias en clases con menos datos.

- **Precisión**

  La precisión mide la proporción de verdaderos positivos sobre el total de instancias predichas como positivas. Aunque en la mayoría de las clases el modelo alcanza una precisión muy cercana al 100%, en algunas clases, como "Apostilla $ 30," la precisión disminuye ligeramente, lo que sugiere que el modelo puede estar menos seguro en la clasificación de instancias de estas clases.

- **Recall**

  El recall mide la proporción de verdaderos positivos sobre el total de instancias que realmente son positivas. La mayoría de las clases mantienen un recall cercano al 100%, pero en clases como "Apostilla $ 30," el recall se reduce al 80%, lo que sugiere que el modelo no logra recuperar todas las instancias positivas de estas clases, posiblemente debido al menor número de muestras.

- **F1-Score:**

  El F1-score, que equilibra precisión y recall, sigue siendo muy alto para la mayoría de las clases. Sin embargo, para clases como "Apostilla $ 30," el F1-score disminuye a 0.89, lo que refleja la ligera pérdida de precisión y recall en esta clase

en particular. Esto sugiere que el modelo podría tener dificultades para mantener un equilibrio perfecto en clases con menos datos.

- **Exactitud Global (Accuracy)**

  La exactitud global del modelo es del 99.68%, lo que sigue indicando que el modelo realiza muy pocas predicciones incorrectas y tiene un rendimiento muy alto en general. No obstante, la ligera variabilidad en precisión y recall en ciertas clases sugiere que el modelo podría beneficiarse de ajustes para mejorar su capacidad en estas áreas específicas.

- **Evaluación del Modelo**

  A pesar del rendimiento excelente del modelo, la ligera caída en precisión y recall en algunas clases podría ser un indicativo de que el modelo enfrenta dificultades para generalizar en clases con menos datos. Es importante continuar evaluando el modelo para asegurarse de que no haya sobreajuste (overfitting) y que pueda generalizar bien a nuevos datos. Validaciones cruzadas adicionales podrían ayudar a identificar cualquier signo de sobreajuste y mejorar la robustez del modelo.

- **Validación Cruzada**

  La validación cruzada sigue siendo una técnica crucial para evaluar la robustez y la capacidad de generalización del modelo Random Forest. En este caso, se utilizó validación cruzada para asegurarse de que el modelo no está sobreajustado a los datos de entrenamiento y puede generalizar bien a nuevos datos. A continuación, se presentan los resultados actualizados y su interpretación.

  - **Proceso de Validación Cruzada**

    El proceso de validación cruzada con k particiones (5-fold cross-validation) fue utilizado para evaluar el modelo de Random Forest. Esta técnica divide los datos en 5 subconjuntos aproximadamente iguales y, en cada iteración, uno de estos subconjuntos se usa como conjunto de prueba mientras los otros 4 se usan para entrenar el modelo. El proceso se repite 5 veces, y los resultados se promedian para obtener una estimación más fiable del rendimiento del modelo:

```python
from sklearn.model_selection import cross_val_score

# Validación cruzada con 5 particiones
cv_scores = cross_val_score(rf_model, X, y_encoded, cv=5)

# Imprimir los resultados
print("Cross-Validation Scores:", cv_scores)
print("Mean Cross-Validation Score:", cv_scores.mean())
print("Standard Deviation of Cross-Validation Scores:", cv_scores.std())
```

Figura 21 Validación cruzada - Random Forest

- **Resultados de la Validación Cruzada**
  - **Cross-Validation Scores:**
    [0.99677939, 0.996139, 0.99327799, 0.997139, 1.0]
  - **Mean Cross-Validation Score:**
    0.9962874762874762
  - **Standard Deviation of Cross-Validation Scores:**
    0.00186314687306748

- **Consistencia del Modelo**

  La baja desviación estándar (0.0019) sugiere que el rendimiento del modelo sigue siendo consistentemente alto en todas las particiones. Esto indica que el modelo es robusto y no es sensible a la forma en que los datos se dividen en las particiones de validación cruzada. Esta consistencia es un buen indicativo de la capacidad del modelo para generalizar bien.

- **Posible Sobreajuste**

  Aunque el rendimiento del modelo es excelente, el riesgo de sobreajuste aún existe, especialmente en clases con pocos datos. La advertencia emitida por sklearn ("**The least populated class in y has only 3 members, which is less than n_splits=5**") sigue siendo relevante. Esto sugiere que algunas clases tienen muy pocas muestras, lo que podría limitar la capacidad del modelo para generalizar adecuadamente en estas clases. Para mitigar este riesgo, sería recomendable realizar una estratificación más rigurosa o aumentar la cantidad de datos en estas clases minoritarias.

- **Robustez del Modelo**

La consistencia en los resultados de validación cruzada, con una desviación estándar baja y un puntaje promedio cercano al 100%, confirma que el modelo Random Forest tiene una buena capacidad de generalización. Sin embargo, para asegurar su robustez en escenarios reales, se recomienda continuar con la validación cruzada y considerar la recolección de más datos en clases minoritarias para mejorar su capacidad de generalización en estas áreas específicas.

4.2.3 Redes Neuronales

Debido a las conclusiones obtenidas con el modelo de regresión lineal y el modelo de Random Forest, se identificó la necesidad de explorar modelos aún más complejos y avanzados para mejorar la precisión de las predicciones.

Una de las técnicas consideradas es el uso de redes neuronales, conocidas por su capacidad para capturar relaciones no lineales complejas y manejar grandes volúmenes de datos. Las redes neuronales son modelos de aprendizaje profundo que consisten en capas de neuronas artificiales que transforman la entrada hasta obtener una salida predictiva.

- **Selección de Variables y División de Datos**
  Al igual que en los modelos anteriores, se seleccionaron las siguientes variables como características predictivas:

Tabla 14 Selección de variables - Redes Neuronales

Características Predictivas
ActoConsularEncoded
MesTramite
AñoTramite
EspecieValor
ValorTramite
DiaTramite
TrimestreTramite
SerieEncoded

La variable objetivo seleccionada fue:

- **EspecieNombre**

- **Entrenamiento del Modelo**

  El conjunto de datos se dividió en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) para evaluar el rendimiento del modelo. Esta división asegura que el modelo se entrena en una porción de los datos y se prueba en datos que no ha visto previamente, lo cual es crucial para evaluar su capacidad de generalización.

```
# Selección de características agrupadas
features = [
    'AñoTramite', 'TrimestreTramite', 'MesTramite', 'DiaTramite',
    'ActoConsularEncoded', 'EspecieValor',
    'ValorTramite', 'SerieEncoded'
]

X = df_grouped[features]
y = df_grouped['EspecieNombre']

# Codificación de la variable objetivo
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# Estandarizar las características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_encoded,
                                                    test_size=0.2,
                                                    random_state=42)
```

Figura 22 Selección de variables y división de datos - Redes Neuronales

- **Definición del Modelo**

  Se definió un modelo de red neuronal con múltiples capas densas (fully connected layers). La arquitectura del modelo incluye una capa de entrada con 256 neuronas, varias capas ocultas con 128 y 64 neuronas, y una capa de salida con una neurona por cada clase de la variable objetivo. Se utilizó la función de activación ReLU (Rectified Linear Unit) para las capas ocultas y la función softmax para la capa de salida, lo que es adecuado para problemas de clasificación multiclase.

```
# Definición del modelo de red neuronal con capas adicionales y Dropout
model = Sequential()
model.add(Dense(256, input_dim=X_train.shape[1], activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(len(le.classes_), activation='softmax'))
```

Figura 23 Definición del modelo - Redes Neuronales

- **Compilación del Modelo**

  El modelo se compiló utilizando la función de pérdida "sparse_categorical_crossentropy" y el optimizador Adam, conocido por su eficiencia en problemas de aprendizaje profundo. La métrica de evaluación seleccionada fue la precisión (accuracy).

```
# Compilación del modelo
model.compile(loss='sparse_categorical_crossentropy',
              optimizer='adam', metrics=['accuracy'])
```

Figura 24 Compilación del modelo - Redes Neuronales

- **Entrenamiento del Modelo**

  El modelo se entrenó con el conjunto de datos de entrenamiento durante 200 épocas, utilizando un tamaño de lote de 64. También se utilizó un conjunto de validación del 20% de los datos de entrenamiento para monitorear el rendimiento del modelo y ajustar los pesos de manera efectiva mediante early stopping.

```
# Ajuste del modelo con early stopping
early_stopping = tf.keras.callbacks.EarlyStopping(monitor='val_loss',
                                                  patience=10,
                                                  restore_best_weights=True)

history = model.fit(X_train, y_train, epochs=200,
                    batch_size=64,  # Ajuste de batch size
                    validation_split=0.2, callbacks=[early_stopping])
```

Figura 25 Entrenamiento del modelo - Redes Neuronales

- **Evaluación del Modelo**

  El rendimiento del modelo se evaluó utilizando la precisión (accuracy) y el reporte de clasificación (classification report). La precisión alcanzó el 99.89% en el conjunto de prueba, lo que indica un rendimiento muy alto. El reporte de clasificación muestra una precisión y recall del 100% para la mayoría de las clases, aunque algunas clases presentan métricas indefinidas debido a la falta de muestras o predicciones correctas en esas clases específicas.

Tabla 15 Muestra de resultados del entrenamiento - Redes Neuronales

Resultados del Entrenamiento:
Epoch 1/200: loss: 1.9666, accuracy: 0.3045, val_loss: 0.9376, val_accuracy: 0.6362
...
Epoch 82/200: loss: 0.0382, accuracy: 0.9892, val_loss: 0.0118, val_accuracy: 0.9987

El historial del entrenamiento muestra una mejora continua en la precisión tanto del conjunto de entrenamiento como del conjunto de validación, alcanzando una precisión muy alta, llegando al 99.89% en el conjunto de prueba. A lo largo de las 82 épocas, el modelo mejoró significativamente su rendimiento, reduciendo la pérdida (loss) y aumentando la precisión en ambos conjuntos.

- **Interpretación de Resultados**
  - **Curvas de perdida y precisión**

    Las curvas de pérdida y precisión durante el entrenamiento y la validación muestran una mejora constante en ambas métricas, alcanzando una precisión del 99.89% en el conjunto de prueba. Este rendimiento indica que el modelo fue capaz de aprender los patrones subyacentes en los datos, pero se debe prestar atención a las clases con menos datos, ya que podrían estar afectando el rendimiento en escenarios con datos nuevos.

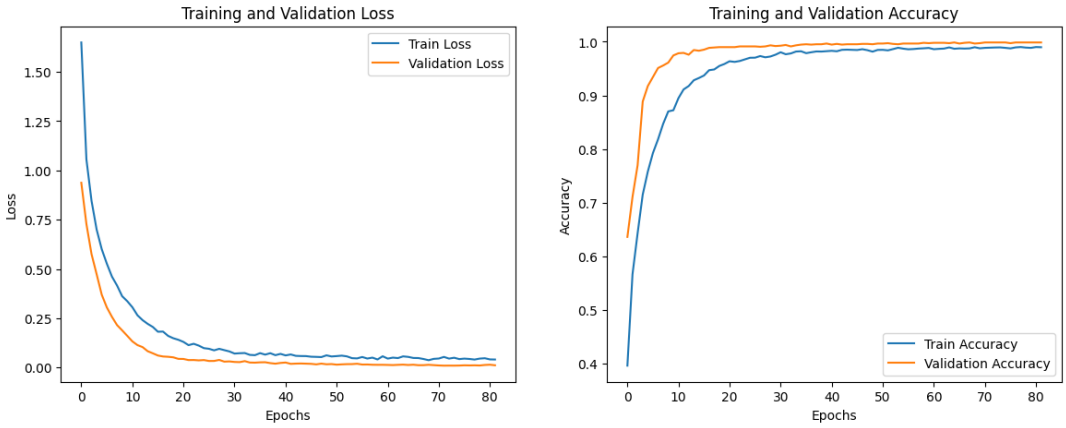

Figura 26 Curva de entrenamiento y validación - Redes Neuronales

- **Matriz de Confusión:**

  La matriz de confusión muestra que el modelo clasificó correctamente la mayoría de las clases, con valores altos en la diagonal que indican Verdaderos Positivos (True

Positives). Sin embargo, algunas clases como "Apostilla ext $40" no tienen suficientes instancias o predicciones correctas, lo que resulta en métricas indefinidas (precisión y recall igual a 0 para esa clase).

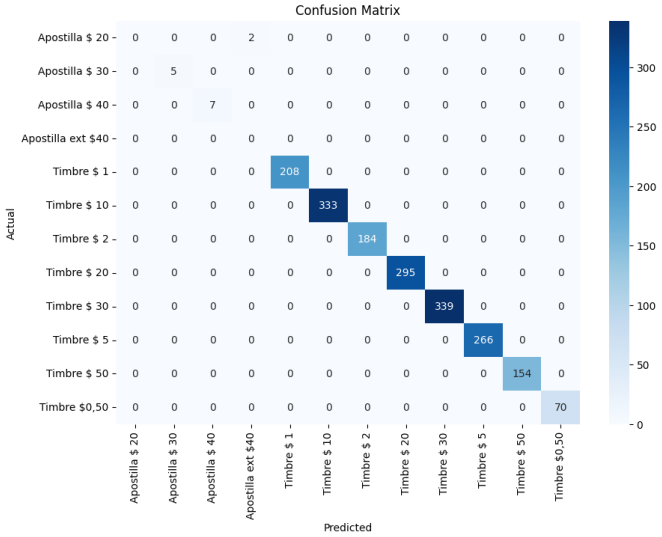

Figura 27 Matriz de confusión - Redes Neuronales

- **Verdaderos Positivos (True Positives):**
  Los verdaderos positivos son altos para la mayoría de las clases, como "Timbre $ 1" y "Timbre $ 10," donde todas las instancias fueron clasificadas correctamente. Sin embargo, algunas clases, como "Apostilla ext $40," no tienen predicciones correctas, lo que puede indicar que el modelo necesita más datos en estas clases minoritarias para mejorar su rendimiento.

- **Falsos Positivos y Falsos Negativos (False Positives y False Negatives):**
  Aunque el modelo tiene un alto rendimiento general, se observan clases con falsos positivos y falsos negativos, como "Apostilla ext $40," donde el modelo no pudo hacer predicciones correctas debido a la falta de instancias. Esto sugiere la necesidad de reentrenar el modelo con más datos en estas clases específicas para mejorar su capacidad de generalización.

  La precisión del 99.89% en el conjunto de prueba es un resultado excepcional. Sin embargo, esta alta precisión puede ser indicativa de sobreajuste (overfitting), especialmente en clases con pocos datos, como "Apostilla ext $40," donde las métricas son indefinidas. El modelo podría haber aprendido demasiado bien los patrones del

conjunto de entrenamiento, lo que afectaría su capacidad para generalizar a datos nuevos.

4.2.4 K-Nearest Neighbors (KNN)

Dado el éxito limitado de los modelos de regresión lineal y la notable precisión del modelo de Random Forest, se decidió explorar otra técnica de aprendizaje automático: K-Nearest Neighbors (KNN). Este algoritmo es conocido por su simplicidad y eficacia en problemas de clasificación y regresión, especialmente cuando se manejan conjuntos de datos de tamaño moderado.

- **Selección de Variables y División de Datos**

  Al igual que en los modelos anteriores, se seleccionaron las siguientes variables predictivas:

Tabla 16 Selección de variables - K-Nearest Neighbors (KNN)

Variables Predictivas
IDTRAMITE
VALORTRAMITE
ACTOCONSULARENCODED
TIPOACTOCONSULAR
ESPECIENOMBREENCODED
SERIEENCODED
NUMERO
MESTRAMITE
AÑOTRAMITE

- **Variable Objetivo**
  - EspecieNombre

El conjunto de datos se dividió en un 70% para el entrenamiento y un 30% para la prueba, manteniendo la consistencia en la evaluación del rendimiento de los modelos.

```python
# Selección de características después del filtrado
features = [
    'ActoConsularEncoded', 'MesTramite', 'AñoTramite', 'EspecieValor',
    'ValorTramite', 'DiaTramite',  'TrimestreTramite',
    'SerieEncoded'
]

X = df_filtered[features]
y = df_filtered['EspecieNombre']

# Codificación de la variable objetivo
le = LabelEncoder()
y_encoded = le.fit_transform(y)

# Estandarizar las características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_encoded,
                                                    test_size=0.3,
                                                    random_state=42)
```

Figura 28 Selección de variables y división de datos - K-Nearest Neighbors (KNN)

- **Entrenamiento del Modelo**

  El modelo KNN fue entrenado utilizando una búsqueda en rejilla (GridSearchCV) para encontrar los mejores hiperparámetros. Los hiperparámetros explorados fueron:

  - **Número de vecinos (n_neighbors):** [3, 5, 7, 9]
  - **Pesos (weights):** ['uniform', 'distance']
  - **Métrica (metric):** ['euclidean', 'manhattan']

  La búsqueda en rejilla determinó que los mejores parámetros fueron:

  - **metric:** 'manhattan'
  - **n_neighbors:** 7
  - **weights:** 'uniform'.

```python
# Definición de la rejilla de hiperparámetros para el GridSearch
param_grid = {
    'n_neighbors': [3, 5, 7, 9],
    'weights': ['uniform', 'distance'],
    'metric': ['euclidean', 'manhattan']
}

# Búsqueda de los mejores parámetros
grid_search = GridSearchCV(knn, param_grid, cv=5)
grid_search.fit(X_train, y_train)
```

Figura 29 Entrenamiento del modelo - K-Nearest Neighbors (KNN)

- **Evaluación del Modelo KNN**

  La evaluación del modelo KNN se realizó utilizando métricas como la precisión (accuracy) y el reporte de clasificación (classification report). El modelo alcanzó una precisión general del 95.56%, lo que representa una mejora significativa en

comparación con versiones anteriores del modelo. A continuación, se detalla el rendimiento en términos de precisión, recall y F1-score para cada clase.

```
# Realizar predicciones
y_pred = knn.predict(X_test)

# Evaluar la precisión
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred, target_names=le.classes_)

print("Accuracy:", accuracy)
print("Classification Report:\n", report)
```

Figura 30 Evaluación del modelo - K-Nearest Neighbors (KNN)

- **Precisión del Modelo**
  - **Accuracy**

    La precisión del modelo fue del 95.56%, lo que indica que el modelo fue correcto en el 95.56% de las predicciones. Este resultado muestra una mejora significativa en la capacidad del modelo para predecir correctamente las clases de especies valoradas.
  - **Reporte de Clasificación**

```
Accuracy: 0.9556191839656406
Classification Report:
                   precision    recall  f1-score   support

   Apostilla $ 20       0.00      0.00      0.00         2
   Apostilla $ 30       1.00      1.00      1.00         5
   Apostilla $ 40       1.00      1.00      1.00        11
Apostilla ext $40       1.00      1.00      1.00         4
       Timbre $ 1       0.93      0.92      0.93       305
      Timbre $ 10       0.93      0.99      0.95       491
       Timbre $ 2       0.90      0.86      0.88       315
      Timbre $ 20       0.98      0.96      0.97       434
      Timbre $ 30       0.97      1.00      0.98       506
       Timbre $ 5       0.98      0.98      0.98       408
      Timbre $ 50       1.00      1.00      1.00       218
     Timbre $0,50       0.95      0.81      0.87        95

         accuracy                           0.96      2794
        macro avg       0.89      0.88      0.88      2794
     weighted avg       0.95      0.96      0.95      2794
```

Figura 31 Resultados del modelo - K-Nearest Neighbors (KNN)

- **Precision, Recall, F1-Score**

  Aunque el modelo muestra un rendimiento excelente en la mayoría de las clases, algunas clases, como "Apostilla $ 20," presentan valores indefinidos debido a la falta

de predicciones correctas. La mayoría de las clases muestran métricas de precisión, recall y F1-score cercanas o iguales al 100%, lo que refleja un rendimiento sólido en la clasificación. Sin embargo, es necesario prestar atención a las clases con menos datos, donde el modelo podría beneficiarse de ajustes adicionales o de la recolección de más muestras.

4.3 Comparación de modelos

Al comparar los resultados obtenidos con los diferentes modelos, se observa una clara diferencia en rendimiento y eficiencia entre ellos.

- **Random Forest**

  Este modelo demostró ser el más robusto y preciso en la mayoría de las clases, con una precisión general cercana al 100%. Aunque no fue el más eficiente en términos computacionales, su capacidad para manejar la variabilidad en los datos y la precisión en la predicción de especies de mayor demanda lo convierten en una opción muy atractiva para el entorno consular. La baja desviación estándar en la validación cruzada también sugiere que el modelo es consistente y confiable.

- **Redes Neuronales**

  Las Redes Neuronales mostraron un rendimiento ligeramente superior al de Random Forest en términos de precisión, alcanzando una precisión de hasta el 99.89% en el conjunto de prueba. Sin embargo, este modelo es significativamente más costoso en términos computacionales y requiere un tiempo de entrenamiento mucho mayor, lo que lo hace menos viable para implementaciones donde la eficiencia operativa es crucial. Además, existe un mayor riesgo de sobreajuste en clases con pocos datos.

- **K-Nearest Neighbors (KNN)**

  El modelo KNN alcanzó una precisión del 95.56%, lo que lo posiciona como una opción decente. Sin embargo, su rendimiento fue inferior al de Random Forest y Redes Neuronales en varias clases, particularmente en aquellas con menos datos. A pesar de su simplicidad y bajo costo computacional, el modelo KNN necesita ajustes adicionales o la incorporación de más datos para mejorar su precisión en clases menos representadas.

- **Regresión Lineal**

  Aunque la regresión lineal fue la opción más sencilla y rápida de entrenar, su precisión fue la más baja de todos los modelos. La regresión lineal tiende a tener dificultades para capturar las relaciones no lineales en los datos, lo que resultó en una capacidad predictiva limitada en comparación con los modelos más complejos.

Tabla 17 Comparación de resultados de modelos.

MODELO	PRECISIÓN (ACCURACY)	PRECISIÓN EN CLASES MENORES	EFICIENCIA OPERATIVA	ADECUACIÓN PARA EL CONSULADO
RANDOM FOREST	99.68%	Alta	Alta	Muy Alta
REDES NEURONALES	99.89%	Media	Baja	Media
K-NEAREST NEIGHBORS	95.56%	Media	Media	Alta
REGRESIÓN LINEAL	Menor	Baja	Muy Alta	Media

De acuerdo a la comparación de la Tabla 17, Random Forest destaca como la opción más adecuada para el consulado debido a su excelente rendimiento y balance entre precisión y eficiencia operativa. Redes Neuronales podría considerarse si se dispone de recursos computacionales adicionales y se prioriza una precisión ligeramente superior. KNN es útil para escenarios donde la simplicidad es clave, pero necesita ser mejorado para igualar a Random Forest en términos de precisión. Regresión Lineal, aunque rápida y eficiente, no es recomendable debido a su baja capacidad predictiva en este contexto.

5. Capítulo V: Conclusiones y Recomendaciones

5.1 Conclusiones

El análisis predictivo de la demanda de especies valoradas en el Consulado del Ecuador en Queens ha representado un paso importante hacia la modernización y optimización de la gestión de recursos. A través del análisis en el desarrollo de modelos de aprendizaje automático, se han obtenido avances significativos en la capacidad de prever la demanda futura, lo que potencialmente permitirá una planificación más eficiente y una mejor asignación de recursos.

Sin embargo, los resultados obtenidos han revelado áreas que requieren mejoras adicionales. Aunque se han logrado niveles notables de precisión en algunos modelos, como Random Forest y Redes Neuronales, otros modelos, como K-Nearest Neighbors (KNN) y Regresión Lineal, no han alcanzado la misma efectividad, especialmente en la predicción de especies con alta demanda. Además, la variabilidad en las predicciones y los desafíos asociados al sobreajuste sugieren que se deben realizar ajustes y optimizaciones continuas para garantizar que los modelos puedan generalizar correctamente y ofrecer predicciones fiables en diversos escenarios.

A continuación, se detallan las conclusiones basadas en los objetivos planteados:

- **Identificación de Trámites Consulares y Demanda**

  Se identificaron con éxito los trámites consulares más frecuentes y su correspondiente demanda de especies valoradas. Este análisis proporcionó información valiosa sobre cómo se distribuye la demanda total anual y mensual, permitiendo una mejor comprensión de los patrones de uso y facilitando la construcción de modelos predictivos.

- **Análisis de Patrones Históricos**

  El análisis de los patrones históricos de uso de especies valoradas durante los años 2021, 2022 y 2023 reveló tendencias clave fundamentales para la predicción futura. Estos datos fueron esenciales para construir modelos que deberán capturar adecuadamente las variaciones temporales y los comportamientos específicos de cada tipo de trámite consular.

- **Desarrollo de Modelos Predictivos**

  Se desarrollaron varios modelos predictivos, incluidos Random Forest, redes neuronales, K-Nearest Neighbors (KNN) y regresión lineal. Random Forest y Redes Neuronales fueron los modelos más precisos, con predicciones que se acercaron al valor real en la mayoría de las especies valoradas. Sin embargo, KNN y Regresión Lineal mostraron limitaciones en la capacidad predictiva, especialmente en especies con demanda más alta.

- **Evaluación de la Efectividad de los Modelos**

  La evaluación de los modelos mediante métricas como el MSE (Error Cuadrático Medio) y el coeficiente de determinación (R^2) mostró que Random Forest y Redes Neuronales son modelos robustos y consistentes. No obstante, los resultados revelan que ninguno de los modelos logra capturar completamente la complejidad de la demanda de especies con alta variabilidad, lo que sugiere la necesidad de ajustes adicionales.

- **Evaluación de Predicciones para el Año 2024**

  Las predicciones para el periodo de enero a junio de 2024 muestran variaciones significativas entre los modelos evaluados. Random Forest predijo con mayor precisión las especies de alta demanda, mientras que Redes Neuronales también mostró un rendimiento competitivo, aunque con variaciones en algunas especies. KNN y Regresión Lineal subestimaron drásticamente los valores reales en muchas especies, lo que resalta la necesidad de seguir ajustando estos modelos:

Tabla 18 Comparación de predicciones con modelos generados

Especie Nombre	TotalReal	Prediccion_KNN	Prediccion_Red_Neuronal	Prediccion_Random_Forest	Prediccion_Regresion_Lineal
Apostilla $ 30	21	141	16	33	110.77
Timbre $ 1	2216	264	260	280	360.89
Timbre $ 2	101	224	228	222	242.04
Timbre $ 20	7109	3652	3960	3960	3518.11
Timbre $ 30	1059	1758	1453	1449	1783.01
Timbre $ 5	3521	3226	3932	4100	3580.55
Timbre $ 50	480	737	1205	1071	880.99
Timbre $0,50	99	229	225	254	309.31

De acuerdo con la comparación de la Tabla 18, Random Forest se mantiene como el modelo más equilibrado en términos de precisión y generalización, mientras que Redes Neuronales ofrece una alternativa competitiva. KNN y Regresión Lineal presentan un desempeño insuficiente para estimaciones precisas en el contexto del consulado.

- **Limitaciones**

  A pesar de los avances, el estudio presenta limitaciones importantes. La cantidad limitada de datos para algunas especies afectó negativamente la capacidad predictiva de los modelos. Además, los modelos más complejos, como las redes neuronales, enfrentaron problemas de sobreajuste, lo que pone de manifiesto la necesidad de mejorar la regularización y la selección de características.

- **Recomendaciones para trabajos futuros**

  Futuros estudios deberían considerar la incorporación de más datos de otros consulados para mejorar la generalización de los modelos. Además, sería beneficioso explorar técnicas de regularización más avanzadas y métodos de reducción de dimensionalidad para mitigar el sobreajuste en modelos complejos.

5.2 Recomendaciones

- Se recomienda mantener un proceso de validación continua y monitoreo regular de los modelos para asegurar que sigan siendo precisos y robustos en diferentes contextos y condiciones. Esto ayudará a identificar y corregir cualquier desviación en el rendimiento del modelo a lo largo del tiempo.

- Se recomienda considerar el uso de técnicas de aprendizaje automático más avanzadas, como redes neuronales profundas y técnicas de ensamblado (ensemble learning), para capturar patrones aún más complejos y mejorar la precisión de las predicciones.

- Se recomienda realizar un análisis exhaustivo de las características más influyentes en las predicciones podría proporcionar información valiosa para mejorar los modelos. Esto permitirá ajustar las variables más críticas y refinar los modelos para mejorar su desempeño.

- Se recomienda reevaluar las características utilizadas, ajustar los hiperparámetros y aplicar técnicas más avanzadas de preprocesamiento y regularización es crucial para mejorar la precisión de las predicciones y hacerlas más útiles para la gestión de stock.

Antes de considerar la implementación de los modelos predictivos en un entorno de producción, es fundamental asegurarse de que estos sean lo suficientemente precisos y confiables para respaldar las decisiones críticas de la gestión del stock en el Consulado del Ecuador en Queens.

Bibliografía

Ministerio de Relaciones Exteriores y Movilidad. (septiembre de 2020). *Arancel Consular y Diplomático [PDF]*. Obtenido de https://www.cancilleria.gob.ec/uploads/2020/09/Arancel-Consular-y-Diplomatico.pdf

Caballero, J. (2020). Uso de modelos predictivos en la gestión tributaria: Un análisis de caso en Latinoamérica. Revista de Economía y Finanzas, 45(2), 123-137.

Fernández, M., & Ruiz, A. (2020). Modelos predictivos en la administración de justicia: Mejora de la eficiencia en la asignación de recursos. Revista de Derecho y Sociedad, 32(1), 45-59.

García, L., & Martínez, S. (2021). Aplicaciones del análisis predictivo en la salud pública: Prevención y control de enfermedades infecciosas. Journal of Public Health Research, 29(3), 203-214.

González, P. (2019). Optimización de procesos en la administración pública mediante modelos predictivos. Revista de Administración Pública, 54(4), 567-590.

López, C., & Ramírez, J. (2018). Predicción de la demanda de visas y pasaportes en consulados: Un estudio de caso. Journal of International Relations and Diplomacy, 12(2), 89-103.

Mendoza, R., Pérez, T., & Silva, M. (2020). Eficiencia en la gestión de citas consulares a través de técnicas de análisis predictivo. Revista de Tecnología y Sociedad, 17(3), 321-335.

Rodríguez, A., & Pérez, F. (2019). Análisis predictivo para la detección de fraude fiscal: Un enfoque basado en big data. Revista de Ciencias Económicas, 34(1), 78-92.

Sánchez, D. (2018). Modelos predictivos en la planificación educativa: Caso de estudio en la asignación de recursos docentes. Journal of Education Research, 21(4), 245-260.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons.

Mitchell, T. M. (1997). ML. McGraw-Hill.

Jordan, M. I., & Mitchell, T. M. (2015). ML: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Breiman, L. (2001). Random Forests. ML, 45(1), 5-32.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc.

García, S. (2018). Introducción a los Algoritmos de ML. Revista Iberoamericana de Inteligencia Artificial, 21(62), 45-56.

Hernández, J., & López, M. (2019). Aplicaciones de Redes Neuronales en la Administración Pública. Revista de Ciencia de Datos y Gobierno, 2(1), 12-29.

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

Breiman, L. (2001). Random Forests. ML, 45(1), 5-32.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.