



**UNIVERSIDAD CÉSAR VALLEJO**

**FACULTAD DE INGENIERÍA Y ARQUITECTURA  
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS**

**Propuesta de un Modelo predictivo para el Tráfico aéreo de  
pasajeros en el Perú, 2023**

**TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:  
Ingeniero de Sistemas**

**AUTOR:**

Pilco Mamani, Jhony Antonio ([orcid.org/0000-0001-9581-6691](https://orcid.org/0000-0001-9581-6691))

**ASESOR:**

Dr. Agreda Gamboa, Everson David ([orcid.org/0000-0003-1252-9692](https://orcid.org/0000-0003-1252-9692))

**LÍNEA DE INVESTIGACIÓN:**

Sistemas de Información y Comunicaciones

**LÍNEA DE RESPONSABILIDAD SOCIAL UNIVERSITARIA:**

Desarrollo económico, empleo y emprendimiento

LIMA – PERÚ

2023

## **Dedicatoria**

*A mis padres Antonio Pilco y Primitiva Mamani, por darme la oportunidad de seguir una carrera en este campo de ingeniería de sistemas y por ser mi inspiración y fuerza motriz.*

*A mis hermanas por su inquebrantable ayuda.*

Jhony Antonio

## Agradecimiento

*Mis mayores fuentes de gratitud son Dios, que nos dio la capacidad de vivir y seguir adelante.*

*A la Universidad César Vallejo, que me permitió completar una meta importante en mi desarrollo profesional, y mi asesor, que con paciencia y pericia nos guio en cada etapa de la preparación de la tesis.*

El autor

## Índice de contenidos

	Pág.
Carátula .....	i
Dedicatoria .....	ii
Agradecimiento .....	iii
Índice de contenidos .....	iv
Índice de tablas .....	v
Índice de figuras .....	vi
Resumen .....	vii
Abstract .....	viii
I. INTRODUCCIÓN.....	1
II. MARCO TEÓRICO.....	4
III. METODOLOGÍA.....	13
3.1 Tipo y diseño de investigación .....	13
3.2 Variables y operacionalización .....	13
3.3 Población, muestra y muestreo .....	14
3.4 Técnicas e instrumentos de recolección de datos.....	15
3.5 Procedimientos.....	15
3.6 Método de análisis de datos.....	17
3.7 Aspectos éticos .....	17
IV. RESULTADOS.....	19
V. DISCUSIÓN.....	28
VI. CONCLUSIONES.....	30
VII. RECOMENDACIONES .....	31
REFERENCIAS.....	32
ANEXOS .....	37

## Índice de tablas

	Pág.
<b>Tabla 1.</b> Análisis descriptivo del tiempo de ejecución de algoritmos .....	19
<b>Tabla 2.</b> Análisis descriptivo del tiempo de ejecución de algoritmos .....	20
<b>Tabla 3.</b> Análisis descriptivo del tiempo de ejecución de algoritmos .....	21
<b>Tabla 4.</b> Estadístico inferencial del tiempo de ejecución de algoritmos .....	23
<b>Tabla 5.</b> Estadístico inferencial de precisión del modelo predictivo .....	23
<b>Tabla 6.</b> Estadístico inferencial de error porcentual absoluto medio.....	23
<b>Tabla 7.</b> Prueba de hipótesis del tiempo de ejecución de los algoritmos.....	25
<b>Tabla 8.</b> Prueba de hipótesis de coeficiente de determinación.....	26
<b>Tabla 9.</b> Prueba de hipótesis del error porcentual absoluto.....	26
<b>Tabla 10:</b> Tabla de Regresión de TheilSen ajuste de parámetro .....	47
<b>Tabla 11:</b> Tabla de predicción de prueba/rechazo.....	47
<b>Table 12:</b> Tabla regresión lineal ajuste de parámetros.....	48
<b>Table 13:</b> Tabla regresión lineal de predicción de prueba/rechazo .....	48
<b>Table 14:</b> Tabla regresión K-nn - ajuste de parámetros .....	49
<b>Table 15:</b> Tabla regresión K-nn de predicción de prueba/rechazo .....	49
<b>Table 16:</b> Tabla de Regresión AdaBoost ajuste de parámetros .....	49
<b>Table 17:</b> Tabla de Regresión AdaBoost predicción de prueba/rechazo.....	50
<b>Table 18:</b> Tabla Árbol de decisión - ajuste de parámetros .....	50
<b>Table 19:</b> Tabla Árbol de decisión predicción de prueba/rechazo .....	50
<b>Table 20:</b> Tabla SVM ajuste de parámetros .....	51
<b>Table 21:</b> Tabla SVM - predicción de prueba/rechazo.....	51
<b>Tabla 22:</b> Modelo predictivo de tráfico aéreo.....	52

## Índice de figuras

	Pág.
Figura 1. Gráfico del tiempo de ejecución de los algoritmos .....	20
Figura 2. Gráfico del tiempo de ejecución de los algoritmos .....	21
Figura 3. Gráfico del tiempo de ejecución de los algoritmos .....	22
Figura 4. Curva de aprendizaje modelo de análisis discriminante lineal .....	52
Figura 5. Gráfico de curva del error de predicción .....	53
Figura 6. Gráfico modelo de análisis discriminante lineal .....	53
Figura 7. Correlación de variables .....	54

## Resumen

En el presente trabajo de investigación, tuvo como objetivo proponer un modelo de predictivo para la toma de decisiones del tráfico aéreo de pasajeros en el Perú 2023. El tipo de investigación fue Aplicada y con diseño Preexperimental. La muestra poblacional fue un dataset del año 2019 al 2022. La metodología empleada para realizar el modelo predictivo fue CRIPS-DM, el lenguaje de programación fue PYTHON y el uso de librerías como Pandas, Numpy y Pycaret, etc., así como una base de datos en Excel. Como resultados se obtuvo: Para el Indicador 1 “Predicción del modelo predicativo” el mejor modelo que fue regresión de TheilSen; Para el Indicador 2 “determinar la métrica de regresión lineal” hubo un ajuste del modelo de 0.9965 y Para el Indicador 3 “determinar el error porcentual absoluto medio para establecer el desempeño” con un indicador de desempeño de 5.8791. Como conclusión general, en base a los tres (3) indicadores el mejor modelo que se ajustó para predecir el tráfico aéreo de pasajeros fue el modelo de regresión de TheilSen.

*Palabras clave: modelo predicativo, machine learning, toma de decisiones, tráfico aéreo de pasajeros.*

## **Abstract**

The objective of this research work was to propose a predictive model for decision-making on-air passenger traffic in Peru 2023. The type of research was applied and with a pre-experimental design. The population sample was a dataset from 2019 to 2022. The methodology used for the predictive model was CRIPS-DM, the programming language was PYTHON and the use of libraries such as Pandas, Numpy and Pycaret, etc., as well as an Excel database. As results were obtained: For Indicator 1 "Prediction of the predictive model" the best model was TheilSen regression; For Indicator 2 "determine the linear regression metric" there was a model fit of 0.9965 and For Indicator 3 "determine the mean absolute percentage error to establish performance" with a performance indicator of 5.8791. As a general conclusion, based on the three (3) indicators the best model fit for predicting air passenger traffic was the TheilSen regression model.

*Keywords: predictive model, machine learning, decision making, air passenger traffic.*



## I. INTRODUCCIÓN

El *modelo predictivo machine learning*, es parte de la informática y un pertenece a la inteligencia artificial tiene como primordial objetivo desarrollar técnicas para que aprendan los computadores, predecir, clasificar y reconocer patrones, pronósticos y aplicar patrones descubiertas recientemente a situaciones que no están comprendidas o cubiertas por un diseño inicial a un conjunto de datos (Zakaryan, 2021).

Las aerolíneas son organizaciones dedicadas al *transporte aéreo de pasajeros*. El mundo de las aerolíneas se clasifica en transporte de pasajeros y mercancía de forma regular, y también existen otras empresas que transportan a sus clientes con puntualidad mutuamente acordada (chavez, 2018).

**En el contexto internacional**, se sostiene que Machine Learning y Data Analytics son las tecnologías más salientes. El análisis de datos ayuda a utilizar varias herramientas y técnicas para analizar, estudiar los datos en un poco tiempo, los algoritmos de aprendizaje automático permiten a los modelos de sistematización utilizando los datos. En el sistema que propuso, el conjunto de datos que utilizaron es el de las aerolíneas indias nacionales y se centraron en un aspecto clave que son las cancelaciones de reservas de pasajeros. El sistema que propusieron y se basaron en técnicas de aprendizaje y Análisis de datos para poder extraer catéteres de suma importancia utilizando herramientas como pandas (Ashad, 2019).

También, se dice que las líneas aéreas el propósito de los *viajes de pasajeros* es de (negocio y placer) tiene una relación fuerte fuerza con el precio de pasajes, las aerolíneas de bajo costo hicieron poco para diferenciar entre los pocos tipos de clientes, en el trabajo que realizo el objetivo principal es modelar la diferencia del comportamiento entre dos tipos de pasajeros. En tal sentido se pretende discutir la importancia y necesidad de técnicas de aprendizaje automático (ML) para facilitar la comunicación de los partes interesados del negocio (Samunderu y Farrugia, 2022).

**En el contexto nacional**, se tiene a las *aerolíneas de viaje en el Perú*, que se expandieron a los todos los lugares, la cual brindan servicios de transporte aéreos a nivel nacional e internacional. Según (DGAC - PERÚ, 2022), las

estadísticas más recientes proporcionadas por Dirección General de Aeronáutica Civil - Gobierno del Perú, el tráfico total de pasajeros en los aeropuertos del Perú aumentó 54,5% en 2021 respecto a 2020, pero disminuyó 51,5% respecto a lo observado en 2019. El tráfico de pasajeros nacionales en 2021 fue 66,9% mayor a lo registrado en 2020, mientras que el tráfico de pasajeros extranjeros se incrementó en 14,1% en el mismo tiempo de análisis, acentuando aún más la recuperación. A excepción de Lima, Cusco tuvo el mayor movimiento global de pasajeros a nivel departamental en 2021, con un incremento de 46,9% respecto a 2020. Le siguieron Arequipa (+64,8%), Piura (+53,3%), Loreto (+47,4%) y San Martín (+81,9%).

En los últimos años el estudio del tráfico aéreo del Perú fue creciendo con el tiempo, pero aun así tiene algunas deficiencias (**problemas específicos**) como son: ocasionando la cancelación de vuelos aéreos; mala toma de decisión, pérdida de clientes por falta de información, falta de atención al cliente por mala toma de decisiones, pérdida de clientes por falta de información del tráfico aéreo.

Se declaró el **enunciado del problema**: *General*: ¿De qué manera la propuesta de un modelo predictivo influye en el tráfico aéreo de pasajeros en el Perú en el año 2023? *Específicos*: ¿De qué medida Machine Learning permite predecir el modelo predictivo de tráfico aéreo de pasajeros?; ¿De qué medida permite predecir la precisión de los modelos de Machine Learning para predecir el tráfico aéreo de pasajeros?; ¿De qué medida Machine Learning permite determinar el error absoluto del modelo de tráfico aéreo de pasajeros?

Se presentó la **justificación de la investigación**: *Conveniencia*, con este trabajo de investigación nos permitirá modelar un análisis predictivo, para la toma de decisiones de la cantidad de vuelos aéreos en los aeropuertos del Perú; *Relevancia social*, este trabajo de investigación beneficiará a las aerolíneas del Perú y los viajeros aéreos (Clientes); *Utilidad metodológica*, concierne para las futuras investigaciones de modelos predictivos del tráfico aéreo; *Implicancias prácticas*, permite solucionar la problemática del tráfico aéreo en los aeropuertos del Perú; *Valor teórico*, ayuda a mejorar las teorías basadas en modelo predictivo y tráfico aéreos.

Se formuló **los objetivos**: *General*: Proponer un modelo de predictivo para la toma de decisiones del tráfico aéreo de pasajere en el Perú 2023; *Específicos*: Determinar el tiempo de ejecución de algoritmos de Machine Learning más apropiado para el tráfico aéreo de pasajeros; Determinar la métrica de precisión del modelo predictivo de Machine Learning para permite predecir el mejor modelo del tráfico aéreo de pasajeros; Determinar el error porcentual absoluto medio para establecer el desempeño del modelo predictivo del tráfico aéreo de pasajeros.

Se infirió las **hipótesis**: *General*: “El desarrollo de un modelo predictivo del tráfico aéreo de pasajeros mejora la toma de decisiones en los aeropuertos del Perú en el año 2023”. *Específicas*: Machine Learning permite predecir el tiempo de ejecución de los algoritmos del tráfico aéreo de pasajeros; Machine Learning permite predecir la precisión del modelo predictivo para predecir el tráfico aéreo de pasajeros; Machine Learning permite predecir el error absoluto del modelo del tráfico aéreo de pasajeros.

## II. MARCO TEÓRICO

En el presente estudio, hemos consultado una serie de precedentes de investigación (**antecedentes**) en revistas científicas y trabajos de investigación que nos han proporcionado información sobre investigaciones anteriores como:

Farzin y Hafizi (2022) en su artículo empleó y comparó dos enfoques probabilísticos y de aprendizaje automático para modelar la elección de la hora de salida, incluidas cuatro opciones, pico de la mañana, pico del mediodía, pico de la tarde y horas no pico. Modelos de máquina de vector de soporte probabilístico (PSVM) y logit multinomial (MNL) calibrados con base en los datos de origen destino de Qazvin, y la evaluación y comparación de estos dos modelos se realizó en base a dos aplicaciones para identificar los factores significativos en el tiempo de salida y predecir la hora de salida. Los resultados muestran que el modelo PSVM puede predecir la hora de salida con una precisión del 53,96 % en comparación con el 49,98 % de precisión del modelo MNL. La máxima precisión equilibrada para predecir las opciones pico de la mañana, pico del mediodía y no pico es del 69 %, 53 % y 60 %, respectivamente; obtenido por el modelo PSVM y el modelo MNL predice la opción pico vespertino con una precisión equilibrada de 52% más precisa que PSVM.

Kim y KeunYoo (2022) en su artículo propuso un procedimiento de dos pasos para predecir el flujo de pasajeros del metro al incorporar la información geográfica del análisis de conglomerados a la predicción de series temporales funcionales. Se utiliza un conjunto masivo de datos de transacciones de tarjetas inteligentes para analizar el número diario de pasajeros de cada estación del metro de Seul. Los resultados pronosticados con la cantidad real de pasajeros demostraron que la cantidad pronosticada de pasajeros basada en los resultados de la agrupación es más precisa en contraste con el resultado sin considerar las propiedades regionales. El resultado de la investigación basada en datos puede aplicar para mejorar el plan de servicio del metro y aliviar las enfermedades infecciosas.

Vallejo y Gaucho(2022) en su artículo tuvo el objetivo de estimar el cálculo de demanda de las paradas de buses en base al análisis de las características de la zona usando la regresión lineal con el uso de un software

estadístico, para evitar los conteos manuales que son costosos y tomas mucho tiempo para tomar los modelos lineales se tomó un total de muestra de 20 paradas en base de la ciudad de Riobamba se realizó un conteo manual de ascenso y descenso, y las características de la zona como la presencia de unidades educativas de nivel de locales comerciales, medio y superior, instituciones financieras y mercados . Con el modelo generado se alcanzó un coeficiente de determinación del 59.93% lo que indica que existe un porcentaje de los datos que pueden ser predichos mediante el uso de las características de las zonas tomadas en cuenta en el presente estudio, sin embargo, el incremento de características de las zonas puede mejorar los resultados del modelo.

Wang y Han (2022), en su artículo propusieron un modelo predictivo de Impulso Gradiente para predecir la disposición de los pasajeros a pagar para elegir sus asientos. Los resultados experimentales muestran que el modelo propuesto supera a los 12 modelos de comparación en términos del área bajo curva característica de trabajo del receptor y la puntuación. Además, estudiaron dos muestras típicas para demostrar la fase de toma de decisiones de un árbol de decisión en BCR-LightGBM y aplicaron el modelo de explicación aditiva de Shapley (SHAP) para analizar los factores de influencia importantes para mejorar aún más la interpretabilidad. Llegando a la conclusión de que los valores del cliente, la tarifa del billete y la duración del viaje son tres factores que las aerolíneas deberían tener en cuenta en su servicio de selección de asientos.

Jiang y Zhan(2022), en su artículo propuso un modelo de selección de características para identificar los elementos que afectan a la felicidad de los pasajeros. En primer lugar, se realiza una selección preliminar de características mediante la eliminación recursiva de características basada en el bosque aleatorio. En segundo lugar, basándose en diferentes modelos de clasificación, KNN, regresión logística, bosque aleatorio, Gaussian Naive Bayes y red neuronal BP. Los resultados experimentales muestran que el modelo RF-RFE selecciona un subconjunto de características que contiene 17 variables. En el modelo de predicción de la clasificación, el bosque aleatorio después de la selección de características RF-RFE muestra el mejor rendimiento de clasificación.

Wu y Gao (2021), en su artículo tuvo el objetivo de predecir el volumen de pasajeros salientes en la estación terminal, se comparó con otros modelos de

predicción, es decir, el LSTM general, el modelo ARIMA y el modelo NAR. Los resultados muestran que el Sp-LSTM puede reducir mejor el error cuadrático medio y el error absoluto medio en comparación con los otros modelos, lo que ilustra la efectividad del modelo propuesto.

Gil y Martínez (2021), en su artículo indicaron que, las aerolíneas suelen reservar en exceso sus vuelos porque algunas de las personas que compraron billetes no se presentarán o cancelarán con antelación. Para determinar la probabilidad de que un pasajero (con atributos específicos de pasajero) llegue o no a su vuelo, se suelen utilizar modelos de previsión convencionales. Mediante el uso de modelos de aprendizaje automático, esta investigación sugiere un enfoque para estimar el número de pasajeros que pierden un vuelo teniendo en cuenta las características del vuelo en lugar del pasajero. Este modelo produce una herramienta fácil de usar que permite al área operativa o de ingresos de una aerolínea obtener una estimación del número real de no-shows en un vuelo. En una aerolínea de bajo coste.

Cheng y Sun (2021), en su artículo propusieron un nuevo marco de pronóstico para el flujo de embarque incorporando el mecanismo generativo en modelos de series de tiempo estándar y aprovechando la fuerte regularidad arraigada en el comportamiento de viaje. Al hacerlo, introdujo el flujo de regreso de viajes de descenso anteriores como una nueva covariable, que captura la estructura causal y las dependencias de largo alcance en los datos de flujo de pasajeros en función del comportamiento de viaje. Desarrollamos el paralelogramo de probabilidad de retorno (RPP) para resumir las relaciones causales y estimar el flujo de retorno. El marco propuesto se evalúa utilizando datos de flujo de pasajeros del mundo real, y los resultados confirman que el flujo de regreso, una sola covariable, puede mejorar de manera sustancial y consistente varias tareas de pronóstico, incluido el pronóstico de un paso adelante, el pronóstico de varios pasos y el pronóstico bajo eventos especiales. Y el método propuesto es más eficaz para las estaciones de tipo empresarial en las que la mayoría de los pasajeros llegan y regresan en el mismo día. Este estudio puede extenderse a otros modos de transporte,

Leurent y Zhu (2021), en su artículo pretendió combinar dichas HTS para conocer el fenómeno de intermodalidad coche-ferrocarril de la movilidad

individual, con una base de datos de Recogida Automática de Tarifas para inferirlo sobre un conjunto muy amplio de viajes individuales. El enfoque implica tres pasos: la caracterización basada en HTS de los viajes Origen-Destino (O-D); la transformación del conjunto de datos AFC mediante la búsqueda dinámica de rutas y las reglas basadas en los datos de la Especificación General de Tránsito (GTFS), para obtener los viajes del ferrocarril; el aprendizaje automático supervisado del uso de P&R basado en los datos HTS y AFC, considerando tres métodos (Support Vector Machine, Decision Tree y Artificial Neural Network). La aplicación a la región de París - Ile-de-France con los datos (Encuestas de viajes de los hogares) de 2010 y AFC de 2019 reveló tres tipos de viajes intermodales mediante un aprendizaje automático no supervisado.

Garikapati y Hou (2020) tuvo como objetivo cuantificar el ahorro de kilómetros en vacío por viaje que se consigue al proporcionar información sobre la demanda futura de viajes a los conductores de las empresas transnacionales. Se emplea un modelo de aprendizaje automático para predecir la demanda de viajes por hora y por intervalo de 10 minutos. Se utilizaron datos de proveedores de servicios de transporte en Austin y Chengdu. Los resultados muestran que este proceso consigue una reducción del 68% - 82% y del 53% - 60% de kilómetros de trayectoria en vacío para las operaciones de muestra de RideAustin y DiDi Chuxing, respectivamente, bajo el supuesto de una disponibilidad ilimitada de aparcamiento a corto plazo.

Abbasi y Imaran (2020) en su artículo propuso soluciones proactivas basadas en la inteligencia artificial (IA) para optimización casi en tiempo real. Aprovechando los datos históricos de los pasajeros En este trabajo se analizaron los flujos de tráfico de pasajeros basándonos en un marco de Acceso, Salida e Intercambio para apoyar la infraestructura ferroviaria contra la congestión. Este trabajo se centra principalmente en el desarrollo de predicciones de flujo de pasajeros utilizando el aprendizaje automático (ML) junto con un novedoso modelo de codificación que es capaz de manejar el flujo de tráfico de pasajeros basado en tiempo real. Amplias simulaciones espacio-temporales aprovechando modelos realistas de predicción de la movilidad muestran que un marco AEI puede lograr un 91,17% de precisión de predicción junto con capacidades de encriptación seguras y ligeras. Los parámetros de seguridad como el coeficiente

de correlación ( $<0,01$ ), la entropía ( $>7,70$ ), la tasa de cambio de píxeles ( $>99\%$ ) intensidad de cambio media unificada ( $>33$ ), contraste ( $>10$ ), homogeneidad ( $<0,3$ ) y energía ( $<0,01$ ) demuestran la eficacia del esquema de codificación propuesto.

Mapkar y Khan (2019) en su artículo presentó un enfoque para reducir el efecto de la cancelación de reservas en las industrias aéreas. Para mitigar los efectos de las cancelaciones, La identificación de los clientes que podrían cancelar sus reservas con anterioridad puede lograrse aprovechando las técnicas de Machine Learning y Data Analytics. El objetivo es diseñar e implementar un modelo de clasificación de aprendizaje automático que prediga la cancelación de una reserva. Nos esforzamos por lograr este objetivo mediante el análisis de los datos de reserva de billetes de una aerolínea india doméstica con la ayuda de técnicas de análisis de datos para encontrar algunos patrones interesantes en el conjunto de datos.

Ziyu y Zhiyin (2017) en su artículo propuso un modelo del promedio móvil integrado autorregresivo estacional para predecir el flujo de pasajeros. El resultado mostró que, este modelo es aplicable a la predicción a corto plazo del tráfico de pasajeros de salida de la terminal del aeropuerto y el error promedio oscila entre el 1 % y el 3 %. El contraste de los valores señalados y reales del flujo de tráfico de pasajeros es bastante pequeño, lo que indica que el modelo tiene una capacidad de predicción del flujo de tráfico de pasajeros bastante buena.

Asad y Imran (2020) se centró principalmente en el desarrollo de predicciones de flujo de pasajeros mediante el uso de Machine Learning (ML) junto con un novedoso modelo de codificación que es capaz de manejar el flujo de tráfico pesado de pasajeros en tiempo real. Se comparó y comunicó el rendimiento de varios modelos de predicción de flujo basados en ML de pasajeros en el mundo real, obtenidos en el metro de Londres y en el Overground. Las extensas simulaciones espacio-temporales que aprovechan modelos realistas de predicción de la movilidad realistas muestran que un marco AEI puede alcanzar una precisión del 91,17% junto con capacidades de cifrado seguras y ligeras de cifrado seguro y ligero. Parámetros de seguridad como el coeficiente de correlación ( $<0,01$ ) entropía ( $>7,70$ ), tasa de cambio de píxeles



(>99%), intensidad de cambio media unificada (>33), contraste (>10), homogeneidad (<0,3) y energía (<0,01) demuestran la eficacia del esquema de cifrado propuesto.

Del mismo modo, es crucial explorar una serie de **bases teóricas** para comprender mejor el tema de estudio propuesto:

*Modelo predictivo:* Es un conjunto de procesos que usan técnicas de computación de estudio de datos que ayudan a predecir la probabilidad de que ocurra ciertas situaciones previas, también se pueden descubrir ciertas oportunidades de negocio, también es un conjunto de métodos computacionales que se basan en la información sobre posibles acontecimientos futuros, lo que permite a la empresa reconocer posibilidades (Zevallos, 2021). Una de las razones por las que un motor de búsqueda en línea como Google funciona tan bien cada vez que se utiliza para buscar en Internet se debe a un algoritmo de aprendizaje que ha dominado el arte de clasificar los sitios web. Numerosas actividades, como la minería de datos, el procesamiento de imágenes, el análisis predictivo, etc. Requieren el uso de estos algoritmos. La ventaja fundamental del aprendizaje automático es que un algoritmo puede funcionar de forma independiente después de aprender a utilizar los datos (Mahesh Batta, 2018). Una de las ventajas del modelo predictivo es emplear algoritmos para realizar el análisis, y en cuanto tenga más datos, mejores los resultados serán mejor. el análisis se realiza mediante algoritmos, que se utilizan para realizar el análisis (Rojas, 2020).

*DataSet:* Es una colección de datos, se denomina dataset en español, que es un anglicismo de la palabra inglesa. Estos conjuntos de datos suelen organizarse en forma de tabla, con un atributo o variable asignado a cada fila y uno a cada columna. Una instancia del conjunto de datos está representada por cada fila y corresponde a un atributo o variable en la columna correspondiente. Estas cualidades, que a veces se denominan rasgos, se presentan de muchas formas: Existen tres tipos de datos: los cuantitativos, que tienen un valor numérico; los ordinales, que tienen una relación de orden; nominales, que son etiquetas que clasifican los datos en grupos (Alvaro y Saucedo, 2020)

*Pasajeros:* Es un conjunto de personas que son transportados de un lugar a otro por medios de transporte terrestres, aéreos marítimos (REAL ACADEMIA ESPAÑOLA).

De la misma manera, para poder comprender el tema de investigación, es importante revisar los **enfoques conceptuales** como:

*Machine learning:* Son algoritmos y modelos relacionados que emplean los sistemas informáticos para realizar una determinada tarea sin que se les haya enseñado explícitamente se conoce como aprendizaje automático. A varios programas de uso cotidiano que incorporan algoritmos de aprendizaje como Google (Mahesh Batta, 2018).

*Algoritmos clustering:* El principal beneficio es el descubrimiento de modelos consistentes que permiten gestionar el gran volumen de datos de forma inteligente y automática para producir información útil como tributo a la resolución de los problemas encontrados(Puga Reyes, 2020).

*Algoritmo de redes neuronales:* son ampliamente utilizadas en una variedad de gamas de aplicaciones de procesamiento de datos. El hecho de que la mayor parte del trabajo informático sólo será necesario durante el entrenamiento es una de las ventajas fundamentales de las redes neuronales. La red neuronal funciona con bastante rapidez una vez que se le ha enseñado un propósito determinado, lo que permite reconocer rápidamente datos desconocidos(Rubio José, 2016).

*Algoritmos de aprendizaje profundo:* es una técnica para enseñar a los ordenadores a imitar el proceso de aprendizaje que emplean las personas. Este modelo puede aprender a ejecutar tareas de clasificación y lograr una mayor precisión que el desempeño humano. El modelo se entrena utilizando un conjunto etiquetado con varias capas y una arquitectura de red neuronal. A diferencia del aprendizaje automático, este procedimiento no requiere la extracción manual de funciones(González Wong, 2019).

*Python:* es un lenguaje de programación que es sencillo de aprender, sin embargo, puede acceder a la potencia del lenguaje de programación a nivel de sistema cuando es necesario. Aparte de las ventajas del propio lenguaje, Python es especialmente atractivo para las cargas de trabajo en el aprendizaje

automático, ciencia de los datos, y la computación científica debido a la comunidad creada en torno a las herramientas y bibliotecas que son fácilmente accesibles (Patterson y Nolet, 2020).

*Aprendizaje supervisado:* se considera el método utilizado durante la fase de entrenamiento en el que hay una o más salidas previstas para un conjunto de valores de entrada, realizados a lo largo de la fase de entrenamiento como valores de entrada. Los datos de entrada son dos objetos, en los que uno representa los datos de entrada y el otro el resultado (Aguirre Ascona, 2019);

*Aprendizaje no supervisado:* Es semejante al agrupamiento, y este proceso no es capaz de realizar un análisis de contraste, por lo que este agrupamiento no se detectó antes. Si optamos por un enfoque no supervisado para extraer palabras clave en secciones específicas, no lo hará. necesita un corpus de anotación manual o automático (Cruz Torrez, 2022).

*Aprendizaje por refuerzo:* Es mejor algoritmo para resolver problemas basados o representados por MDP es el aprendizaje por refuerzo. Se ha dicho que el aprendizaje por refuerzo implica "hacerlo mal y aprender de ello", lo cual es una noción intuitiva. Más detalladamente, el aprendizaje por refuerzo tiene seis fases principales, y son; Observación de entorno, Toma de decisión, ejecución de la decisión tomada, Asignación de la recompensa, Mejora de nuestra estrategia y Repetición. En términos generales el aprendizaje por refuerzo se puede ejecutar por modelos y sin modelos (Rosario, 2021).

En cuanto a la **metodología de desarrollo**, se conceptualizó tres (3) metodologías de machine learning para el desarrollo del modelo predictivo, donde se eligió Cross-Industry Standard Process for Data Mining (CRIPS-DM) mediante el *método del juicio experto* (ver Anexo 3).

CRIPS-DM: Metodología que se basa en un modelo de proceso jerárquico, descrito en cuatro fases niveles de abstracción: recolección de datos, selección de la información, modelado experimental, evaluación, Deployment. En general, ofrece una visión general de las fases que componen la vida de un proyecto de aprendizaje automático (Ritha y Andre, 2020).

SEMMA: Metodología desarrollada por el instituto SAS que considera un ciclo con 5 etapas para el proceso: Muestrear, explorar, modificar, modelar y

evaluar. El modelado de datos es el proceso de descubrir información predictiva a partir de grandes análisis de base de datos (Rasheed y Gertrude, 2019).

KDD: Metodología para descubrir conocimiento interesante y útil a partir de una base de datos. Es sólo un paso del proceso KDD en el que se aplica un algoritmo para encontrar patrones en los datos. Se centra en el proceso global de descubrimiento de conocimiento a partir de datos, que cómo se almacenan los datos y se accede a ellos, cómo pueden utilizarse los algoritmos pueden utilizarse en conjuntos de datos masivos y ejecutarse de forma eficaz, y cómo interpretar y visualizar los resultados (Dåderman y Rosander, 2018).

### III. METODOLOGÍA

#### 3.1 Tipo y diseño de investigación

- Tipo de investigación

*Aplicada* porque se utilizó el aprendizaje automático para la predicción del tráfico aéreo de pasajeros, con el uso de Machine Learning (Ramos-Galarza, 2021), Se caracteriza por la asignación de probabilidades de los participantes en un grupo de experimental de estudios, la diferentes condiciones que no son controladas por el investigador no se distribuyen al azar.

- Diseño de Investigación

*Preexperimental* debido a que se tomó una base de datos del tráfico aéreo para aplicar Machine Learning y obtener una observación y realizar una medición. (Ramos-Galarza, 2021), los preexperimentales recibe la intervención del investigado que aplica una mediada de una o más variables para ver la variable en un grupo y luego saber cómo le está yendo al grupo.

#### 3.2 Variables y operacionalización

- Variables

- Independiente: Modelo predictivo

- Definición conceptual:

“Conjunto de estrategias que predicen resultados futuros utilizando el aprendizaje automático, la minería de datos históricos, el big data y el reconocimiento de patrones para la mejora de la toma de decisiones mediante técnicas de análisis de datos”(Shenzhen, 2022).

- Definición operacional:

Se puede medir a través de análisis de los datos como: el procesamiento de los datos usando machine learning, para la toma de decisiones.

- Dependiente: Tráfico aéreo

- Definición conceptual:

“Es el movimiento continuo que se produce en el aire. Necesita una supervisión y gestión periódicas a través de una serie de procesos y un estricto conjunto de directrices”(Ferrovial, 2022).

- Definición operacional

Se puede medir por el análisis exploratorio para utilizar un modelo predictivo y analizar los datos y predecir el tráfico aéreo para la toma de decisión.

- Operacionalización

La variable de operacionalización del estudio se encuentra específicamente en la matriz de operacionalización (ver Anexo 2).

### 3.3 Población, muestra y muestreo

- Población

La población se encuentra determinada por operaciones de del tráfico aéreo durante la semana.

$$N = \frac{140 \text{ vuelos}}{\text{dia}} \times \frac{7 \text{ dias}}{\text{semana}} = 980 \text{ vuelos por semana}$$

$$N = 980 \text{ vuelos por semana}$$

- Muestra

La muestra se encuentra determinado por el elemento de la población; esto es debido a que la población es mayor a 30.

$$N = 980 \text{ vuelos a la semana}$$

- Muestreo

De tipo no probabilístico.

### 3.4 Técnicas e instrumentos de recolección de datos

- Técnica:
  - Observación. Se uso esta técnica porque se observó el comportamiento de los algoritmos de machine Learning para el modelo predicativo del tráfico aéreo de pasajeros.
- Instrumentos:
  - Ficha de datos: Es el conjunto de datos donde está registrado el tráfico aéreo de pasajeros.

### 3.5 Procedimientos

A continuación, se explica el enfoque que se siguió para el cumplimiento de los objetivos específicos de la investigación:

- *Objetivo específico 1: Determinar el tiempo de ejecución de algoritmos de Machine Learning más apropiado para el tráfico aéreo de pasajeros*

Para poder determinar el modelo predictivo de tráfico aéreo de pasajeros se utilizó la metodología CRIPS-DM, se deberán de realizar los siguientes pasos: Tener clara la pregunta que queremos responder; Tener una idea general de nuestro dataset; Definir los tipos de datos que tenemos; Elegir el tipo de estadística descriptiva Visualizar los datos; Analizar las posibles interacciones entre las variables del dataset; y finalmente

Extraer algunas conclusiones de todo este análisis, se realizará el análisis usando librerías de Python como pandas, numpy, pycaret, etc. Y con ello determinar la precisión del algoritmo Machine Learning con los datos de estudio,

- *Objetivo específico 2: Determinar la métrica de precisión del modelo predictivo de Machine Learning para permite predecir el mejor modelo del tráfico aéreo de pasajeros*

Para determinar el mejor valor de regresión lineal de los algoritmos de Machine Learning se predicó los resultados del tráfico aéreo de pasajeros. Los 6 modelos comparados, luego se luego se creará modelos para cada algoritmo de Machine Learning para poder obtener un resultado mucho mejor que al primer entrenamiento realizado, y así seguir una serie de pasos que fue descrita en el primer objetivo específico, con el uso de Python y las librerías como pandas, numpy, matplotlib, Pycaret, etc. Las cuáles serán necesarias para poder realizar el modelo de predicción con todas estas librerías empezará a procesar los datos y así buscar un modelo predictivo más exacto, con el entrenamiento realizado el mismo modelo mostrará el resultado más propicio del tráfico aéreo de pasajeros.

- *Objetivo específico 3: determinar el error porcentual absoluto medio para establecer el desempeño del modelo predictivo del tráfico aéreo de pasajeros*

Para poder determinar el porcentaje de error de modelos entrenados del tráfico aéreo con Machine Learning, y para permitirle comparar la precisión de los métodos de series de igual manera se realizó una secuencia de algoritmos para cada modelo como la obtención de datos desde el dataset luego se procesará el datade con una serie de algoritmos de inteligencia artificial de método supervisado con el lenguaje de programación de Python, con ayuda de entrenamiento del algoritmo, dará un resultado que muestre la precisión de la serie de tiempo de los modelos.



### **3.6 Método de análisis de datos**

Se aplicó el método de análisis y procesamiento de datos tanto como nivel descriptivo e inferencial.

*Estadística descriptiva:* Hace sugerencias sobre cómo presentar los datos de la investigación en tablas, figuras, cuadros o gráficos de forma clara y comprensible. Antes de realizar un análisis descriptivo, es fundamental establecer las escalas de medición de las numerosas variables que se registraron en la investigación, así como esbozar el propósito o los objetivos del análisis. Las tablas y los gráficos tienen por objeto presentar información actual sobre los resultados de una investigación. Los gráficos, que pueden ser histogramas, gráficos circulares, gráficos de líneas, gráficos de "caja y bigotes" o gráficos de dispersión, muestran tendencias. Las imágenes se utilizan para ilustrar ideas o hechos. En función de los objetivos del estudio (Rendón, Villasís y Miranda, 2016).

*Estadística inferencial:* Se refiere a un grupo de técnicas estadísticas que permiten inferir correlaciones entre variables y la distribución de la población a partir de los datos suministrados por la muestra utilizada. Por lo tanto, la estimación y la prueba de hipótesis son los objetivos esenciales de la inferencia estadística (Veiga y Torres, 2020).

### **3.7 Aspectos éticos**

El presente trabajo de investigación defiende la exactitud de los resultados de la información suministrada por la División de Gestión de Servicios. La exactitud de los resultados de la División de Gestión de Servicios y la confidencialidad de los datos que ha suministrado. Los autores que figuran en este estudio de investigación defienden sus derechos de propiedad intelectual.

La universidad Cesar Vallejos está ligada a un código de ética con RCU N° 0126-2017/UCV.

El sistema Turnitin ofrece la posibilidad de comprobar los trabajos de los alumnos en busca de errores y posibles plagios comparándolos con una amplia base de datos de fuentes. Informa del porcentaje de un trabajo que es similar o citado de otras fuentes.

El sistema de norma ISO-690 es una norma de la Organización Internacional de Normalización que ofrece líneas fundamentales para la creación de reseñas bibliográficas a obras que han sido publicadas, ya sea impresa o no. No se aplica a las referencias de manuscritos u otras obras que aún no han sido publicadas, ni a las citas de leyes o jurisprudencia (Universidad de Alcalá, 2022).

#### IV. RESULTADOS

- **Estadística descriptiva:**

**indicador 1:** Tiempo de ejecución de los algoritmos

**Tabla 1.** Análisis descriptivo del tiempo de ejecución de algoritmos

<b>Estadísticos descriptivos</b>					
	N	Mínimo	Máximo	Media	Desviación estándar
PRE-Test Tiempo ejecución	6	0,043	0,069	0,05033	0,009543
POST-Test Tiempo ejecución	6	0,046	0,124	0,08900	0,025573
N válido (por lista)	6				

**Fuente:** elaboración propia

Se realizaron la comparación de 6 algoritmos predictivos Machine Learning teniendo los 6 algoritmos del pretest de tiempo de ejecución, se puede apreciar que se tiene un mínimo de 0,043 de acierto y un máximo de 0,069 en los 6 algoritmos el cual obtiene una media de 0,05033 y una desviación de 0,009543, en el posttest de tiempo de ejecución de tiene un mínimo de 0,046 de acierto y un máximo de 0,124 con una media de 0,05033 de acierto.

Podemos apreciar en la figura 1 que, con el entrenamiento de los algoritmos para el modelo predictivo de tráfico aéreo de pajareros con un tiempo de ejecución de 0,089 lo que indica que fue mejorando en el tiempo de ejecución cada vez que entreno el algoritmo de Machine Learning.

Figura 1. Gráfico del tiempo de ejecución de los algoritmos



**Fuente:** Elaborado por el autor en base a los datos obtenidos

**indicador 2:** Precisión del modelo predictivo

**Tabla 2.** Análisis descriptivo del tiempo de ejecución de algoritmos

Estadísticos descriptivos					
	N	Mínimo	Máximo	Media	Desviación estándar
PRE-Test Error porcentual	6	-0,1184	0,9961	0,8038	0,4518
POST-Test Error porcentual	6	-0,0769	0,9965	0,8129	0,4359
N válido (por lista)	6				

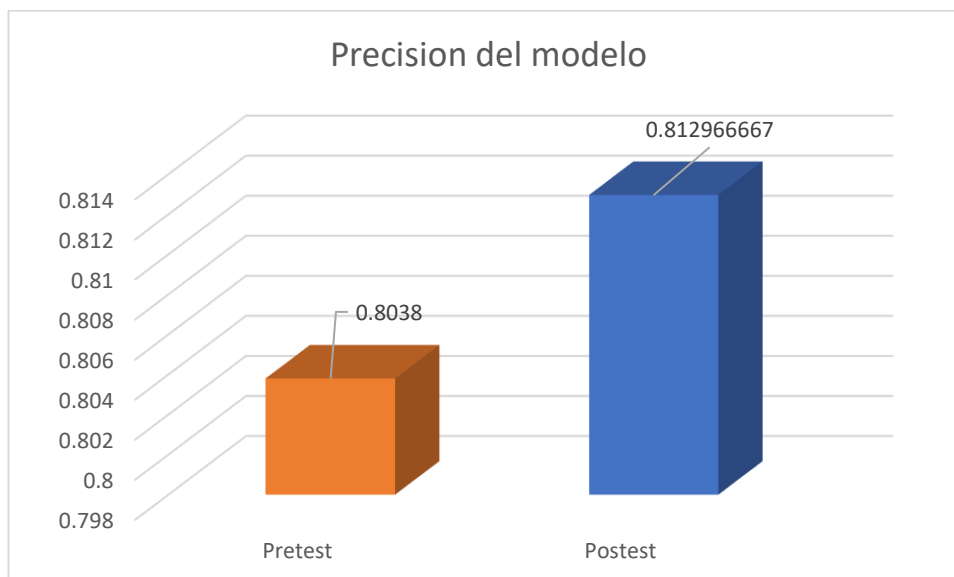
**Fuente:** elaboración propia

Se realizaron la comparación de 6 algoritmos predictivos Machine Learning teniendo los 6 algoritmos del pretest de la precisión del modelo, se puede apreciar que se tiene un mínimo de -0,1184 de acierto y un máximo de 0,9961 en los 6 algoritmos el cual obtiene una media de 0,8038 y una desviación de 0,4518, en el posttest de tiempo de ejecución de tiene un mínimo de -0,0769 de acierto y un máximo de 0,9965 con una media de 0,4359 de acierto y una desviación de 0,4359.

Podemos apreciar en la figura 2 que, la precisión del modelo predictivo del algoritmo de Machine Learning con el entrenamiento de los modelos predictivos

del tráfico aéreo de pajareros, tuvo una precisión del 0.8129, lo que se puede apreciar que en cada entrenamiento que tuvo los modelos obtuvo una precisión más acertada.

Figura 2. Gráfico del tiempo de ejecución de los algoritmos



**Fuente:** Elaborado por el autor en base a los datos interpretados

### indicador 3: Error porcentual absoluto medio

**Tabla 3.** Análisis descriptivo del tiempo de ejecución de algoritmos

Estadísticos descriptivos					
	N	Mínimo	Máximo	Media	Desviación estándar
PRE-Test Error porcentual	6	-24.51	34.54	5.55	19.09
POST-Test Error porcentual	6	3.41	58791.00	9809.17	23996.09
N válido (por lista)	6				

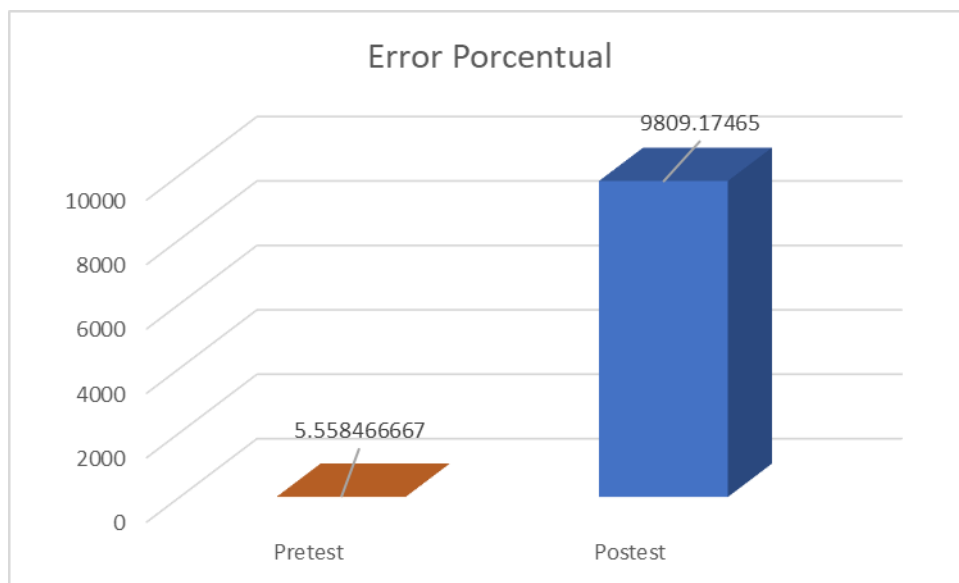
**Fuente:** elaboración propia

Se realizaron la comparación de 6 algoritmos predictivos Machine Learning teniendo los 6 algoritmos en pretest en el error porcentual absoluto, se puede apreciar que se tiene un mínimo de -24.51 de acierto y un máximo de 34.54 en los 6 algoritmos en el cual se obtiene una media de 5.55 y una desviación de 19.09, en el posttest de tiempo de ejecución de tiene un mínimo de 3.41 de acierto

y un máximo de 58791.00 con una media de 9809.17 de acierto y una desviación de 23996.09.

Podemos apreciar en la figura 3 que, el error porcentual absoluto medio del modelo predictivo del algoritmo de Machine Learning, con el entrenamiento de los modelos del tráfico aéreo de pajareros, tuvo un error porcentual absoluto medio del 9809.17, se puede apreciar que, en los entrenamientos de los modelos, el porcentaje de error tuvo un aumento de validación de datos en el posttest.

Figura 3. Gráfico del tiempo de ejecución de los algoritmos



**Fuente:** Elaborado por el autor en base a los datos interpretados

- **Estadística Inferencial:**

**Indicador 1:** Tiempo de ejecución de los algoritmos

**Tabla 4.** Estadístico inferencial del tiempo de ejecución de algoritmos

	<b>Pruebas de normalidad</b>					
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
PRE-Test Tiempo ejecución	0,305	6	0,084	0,758	6	0,024
POST-Test Tiempo ejecución	0,256	6	0,200*	0,938	6	0,643

**Fuente:** los datos fueron obtenidos aplicando SPSS

Como se muestra en la tabla anterior los resultados de la prueba indica que uno de los valores Sig. Es  $< 0.05$  por lo tanto los datos no tienen una distribución normal y se aplicara la prueba no paramétrica de Wilcoxon.

**Indicador 2:** Precisión del modelo predictivo

**Tabla 5.** Estadístico inferencial de precisión del modelo predictivo

	<b>Pruebas de normalidad</b>					
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
PRE-Test Error porcentual	0,487	6	0,000	0,508	6	0,023
POST-Test Error porcentual	0,484	6	0,000	0,508	6	0,002

**Fuente:** elaboración propia

Como se muestra en la tabla anterior los resultados de la prueba normalidad de precisión del modelo predictivo indica que uno de los valores Sig. Es  $< 0.05$  por lo tanto los datos no tienen una distribución normal y se aplicara la prueba no paramétrica de Wilcoxon.

**Indicador 3:** Error porcentual absoluto medio

**Tabla 6.** Estadístico inferencial de error porcentual absoluto medio

	<b>Pruebas de normalidad</b>					
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
PRE-Test Error porcentual	0,257	6	0,200*	0,940	6	0,656
POST-Test Error porcentual	0,492	6	0,000	0,496	6	0,001

**Fuente:** elaboración propia

Como se muestra en la tabla anterior los resultados de la prueba de normalidad del error porcentual absoluto medio indica que uno de los valores Sig. Es  $< 0.05$  por lo tanto los datos no tienen una distribución normal y se aplicara la prueba no paramétrica de Wilcoxon.



- **Prueba de hipótesis:**

**Nula (H0):** El Modelo Predictivo del Tráfico aéreo de Pasajeros no mejorara la toma de decisiones en los Aeropuertos del Perú, en el año 2023

**Alternativa (Ha):** El Modelo Predictivo del Tráfico aéreo de Pasajeros mejorara la toma de decisiones en los Aeropuertos del Perú, en el año 2023

**indicador 1:** Tiempo de ejecución de los algoritmos

Hipótesis (H1) Machine Learning permite predecir el tiempo de ejecución de los algoritmos del tráfico aéreo de pasajeros.

**Tabla 7.** Prueba de hipótesis del tiempo de ejecución de los algoritmos

<b>Estadísticos de prueba<sup>a</sup></b>	
	POST-Test
	Tiempo
	ejecución - PRE-
	Test Tiempo
	ejecución
Z	-2,201 <sup>b</sup>
Sig. asin. (bilateral)	0,028

**Fuente:** elaboración propia

**Resultado:**

El valor obtenido del Sig. (0,028) < 0.05, significa que rechaza a la hipótesis nula y acepta a la hipótesis alterna (Ha) que indica: El Modelo Predictivo del Tráfico aéreo de Pasajeros mejorara el tiempo de ejecución para el tráfico aéreo de pasajeros del Perú, en el año 2023

## Indicador 2: Coeficiente de determinación

Hipótesis (H2) Machine Learning permite predecir la precisión del modelo predictivo para predecir el tráfico aéreo de pasajeros.

**Tabla 8.** Prueba de hipótesis de coeficiente de determinación

Estadísticos de prueba <sup>a</sup>	
	POST-Test Error porcentual - PRE-Test Error porcentual
Z	-1,572 <sup>b</sup>
Sig. asin. (bilateral)	0,016

**Fuente:** elaboración propia

El valor obtenido del Sig. (0,016) < 0.05, significa que rechaza la hipótesis nula y acepta la hipótesis alterna (H<sub>a</sub>) que indica: que Machine Learning permite predecir la métrica de precisión del modelo predictivo para el tráfico aéreo de pasajeros.

## Indicador 3: Error porcentual absoluto medio

Hipótesis (H3) Machine Learning permite predecir el error absoluto del modelo del tráfico aéreo de pasajeros.

**Tabla 9.** Prueba de hipótesis del error porcentual absoluto

Estadísticos de prueba <sup>a</sup>	
	POST-Test Error porcentual - PRE-Test Error porcentual
Z	-1,363 <sup>b</sup>
Sig. asin. (bilateral)	0,023

**Fuente:** elaboración propia

El valor obtenido del Sig. (0,023) < 0.05, significa que rechaza la hipótesis nula y acepta la hipótesis alterna (Ha) e indica: Machine Learning permite predecir el error porcentual absoluto del modelo predictivo del tráfico aéreo de pasajeros.

## V. DISCUSIÓN

A continuación, se exponen las discusiones en base a los resultados obtenidos durante el estudio de la investigación.

Con referencia al indicador 1: “predicción para toma de decisión”, Para la predicción del modelo se utilizó la metodología de CRISP-DM, la cual guio cada fase del modelado y la realización de cada tarea, para completar el proyecto. El uso de la metodología utilizada en este trabajo se puede corroborar en los resultados Sergio García Alonso(2021), que en sus conclusiones afirma que la utilización de la metodología CRISP-DM ha guiado en las fases y tareas realizadas en cada etapa de la evaluación del modelo predictivo, para la predicción de modelos Machine Learning, la metodología resulta ser ágil a la hora de realizar cada tarea del modelo predictivo Ansari Ahlam, Shaikh Ashad(2019), en su conclusión afirma que analizó el conjunto de datos de una aerolínea india y aplicó una serie de pasos de CRISP-DM, comprensión del negocio, comprensión de los datos, preparación de los datos, modelado evaluación y despliegue. Para entrenar diferentes modelos de aprendizaje que se compararon.

Con referencia al indicador 2: “determinar la métrica de regresión lineal”, los resultados obtenidos en la primera etapa de la evolución de los modelos se muestran en la tabla 19, del modelo de regresión de TheilSen teniendo el resultado de la métrica de regresión lineal  $R^2$  0.9961 a diferencia de los otros modelos comparados que también tienen un buen resultado, Fosaca Gamarra Almudena (2020), afirma que el mejor modelo SVM con una métrica de regresión lineal  $R^2$ , del modelo que realizó en comparación de dos algoritmos comparados y exploración, optimización de estos modelos de Machine Learning con un resultado de optimización de 0.807 a 0.97, Darwin Giuseppe(2019), afirma que el estudio de la investigación requirió el uso de características, es decir que a mayor uso de métricas se obtuvieron mejores resultados cuadráticos, se estableció una métrica para poder obtener un mejor resultado cuadrático, Han J(2022), en su trabajo de investigación concluye que el uso del modelo de Regresión de TheilSen tuvo mucho un resultado óptimo

a la hora de entrenar los modelos teniendo un resultado de  $R^2$  de 0.6845, lo que indica que el modelo es óptimo para el análisis de datos.

Con referencia al indicador 3: “determinar el error porcentual absoluto medio para establecer el desempeño”, los resultados obtenidos en el entramiento del modelo de regresión de TheilSen con una regresión lineal  $R^2$  de 0.9964 y un error absoluto medio  $MAPE$  de 5.8791, lo cual nos indica un buen desempeño del pronóstico del modelo del tráfico aéreo de pasajeros a diferencia del demás modelo que se observan en los resultados comparados. Darwin Giuseppe(2019), indica que se validaron los modelos con el uso de métricas y su correspondiente indicador de error de (RMSE) con el que se tuvo resultados de demostración, obteniendo el error mínimo inferir que el modelo usado es la indicada, Zuñiga Quiñones(2021), afirma que la visualización y comparación de los resultados de se observa el valor del error cuadrático  $R^2$  de cada uno de los modelos se seleccionó el error cuadrático por que a diferencia de del error medio Absoluto en su magnitud se otorga relativamente alto a los errores grandes, lo que se consideraría significativamente al contexto del modelo predicativo.

## VI. CONCLUSIONES

1. Se logró concluir que, después de entrenar varios algoritmos como, regresión lineal, Regresión TheilSen, K-nn, Regresión AdaBoost, árbol de decisión y Máquina de vectores de soporte, el mejor modelo predictivo de Machine Learning que muestra los mejores datos de precisión en el modelado predictivo para el tráfico aéreo el modelo de regresión de TheilSen con un tiempo de ejecución de 0.084 lo que indica es un valor cercano al valor uno y es óptimo para poder realizar el modelo predicativo.
2. Se logró concluir que, luego de realizar varios entrenamientos de algoritmos de Machine Learning, Regresión de TheilSen, K-nn, Regresión AdaBoost, árbol de decisión y Máquina de vectores de soporte, que el modelo de Machine Learning con más precisión del modelo predictivo es regresión Thiessen con un resultado de 0.9965 brinda un mejor resultado a la predicción del entrenamiento del modelo de tráfico aéreo de pasajeros en comparación a los otros modelos entrenados.
3. Se logró concluir que, usar varios algoritmos de Machine Learning, TheilSen, K-nn, Regresión AdaBoost, árbol de decisión y Máquina de vectores de soporte, que el modelo de Machine Learning para el modelo se requirió uso ajustes de hiperparámetros del algoritmo de Machine Learning, para obtener mejores resultados un error absoluto medio *MAPE*, para que el modelo se ajuste mejor a la realidad.

## **VII. RECOMENDACIONES**

A los responsables del tráfico aéreo del país:

Se recomienda aplicar otros algoritmos de Machine Learning a los datos de tráfico aéreo de pasajeros, con el propósito de aumentar la perspectiva de predicciones, con datos más exactos y completos ya que el dataset tuvo una pérdida de datos a la hora de preparar el dataset para el modelado.

A los responsables de informática:

Se recomienda realizar más comparaciones con otros algoritmos de Machine Learning, para ver si existe otro algoritmo más adecuado para el modelo predictivo.

A los directivos del transporte aéreo del país:

Se recomienda crear un modelo predictivo empleando más factores relacionados al tráfico de pasajeros, como venta de pasajes, satisfacción de pasajeros, etc.

A los interesados:

Se recomienda trabajar más los datos de tráfico aéreo de pasajeros para poder tener un modelo predictivo mucho más exacto.

## REFERENCIAS

- Aguirre Ascona. (2019). *UNIVERSIDAD PERUANA UNIÓN FACULTAD DE INGENIERÍA Y ARQUITECTURA Escuela Profesional de Ingeniería de Sistemas*.  
[https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/2511/Aguirre\\_Trabajo\\_Bachillerato\\_2019.pdf?sequence=4&isAllowed=y](https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/2511/Aguirre_Trabajo_Bachillerato_2019.pdf?sequence=4&isAllowed=y)
- Alvaro, J., & Saucedo, P. (2020). *MODELO PREDICTIVO PARA EL RECONOCIMIENTO DE ACTIVIDADES*.
- Ansari Ahlam, Khan, M., Shaikh Ashad, & Mapkar Salim. (2019). *Cancellation Prediction for Flight Data Using Machine Learning*. <http://ssrn.com/link/2019-ICAST.html>
- Ansari Ashad, Shaikh Salim, Mapkar, A., & Khan, M. (2019). *Cancellation Prediction for Flight Data Using Machine Learning*. <http://ssrn.com/link/2019-ICAST.html>
- Aquino chavez, F. M. (2018). *CARACTERIZACIÓN TRIBUTARIA DEL TRANSPORTE AÉREO DE PASAJEROS EN EL PERÚ*.  
[https://repositorio.uladech.edu.pe/bitstream/handle/20.500.13032/7346/CARACTERIZACION\\_TRIBUTARIA\\_TRANSPORTES\\_AEREOS\\_DE\\_PASAJEROS\\_AQUINO\\_CHAVEZ\\_FANNY\\_MARITZA..pdf?sequence=4&isAllowed=y](https://repositorio.uladech.edu.pe/bitstream/handle/20.500.13032/7346/CARACTERIZACION_TRIBUTARIA_TRANSPORTES_AEREOS_DE_PASAJEROS_AQUINO_CHAVEZ_FANNY_MARITZA..pdf?sequence=4&isAllowed=y)
- Asad, S. M., Ahmad, J., Hussain, S., Zoha, A., Abbasi, Q. H., & Imran, M. A. (2020a). Mobility prediction-based optimisation and encryption of passenger traffic-flows using machine learning. *Sensors (Switzerland)*, 20(9).  
<https://doi.org/10.3390/s20092629>
- Asad, S. M., Ahmad, J., Hussain, S., Zoha, A., Abbasi, Q. H., & Imran, M. A. (2020b). Mobility prediction-based optimisation and encryption of passenger traffic-flows using machine learning. *Sensors (Switzerland)*, 20(9).  
<https://doi.org/10.3390/s20092629>
- Cheng, Z., Trépanier, M., & Sun, L. (2021). Incorporating travel behavior regularity into passenger flow forecasting. *Transportation Research Part C: Emerging Technologies*, 128. <https://doi.org/10.1016/j.trc.2021.103200>



- Cruz Torrez. (2022). *ALGORITMOS DE APRENDIZAJE AUTOMÁTICO NO SUPERVISADO PARA LA EXTRACCIÓN DE PALABRAS CLAVES EN TRABAJOS DE INVESTIGACIÓN DE PREGRADO*.  
[http://repositorio.unap.edu.pe/bitstream/handle/UNAP/18372/Fred\\_Torres\\_Cruz.pdf?sequence=1&isAllowed=y](http://repositorio.unap.edu.pe/bitstream/handle/UNAP/18372/Fred_Torres_Cruz.pdf?sequence=1&isAllowed=y)
- Dåderman, A., & Rosander, S. (2018). *Evaluating Frameworks for Implementing Machine Learning in Signal Processing A Comparative Study of CRISP-DM, SEMMA and KDD*.
- Darwin Giuseppe Marín Vilca, & Ian Augusto Pineda Torres. (2019). "Modelo predictivo Machine Learning aplicado a análisis de datos Hidrometeorológicos para un SAT en Represas."
- David, J., Gil, C., & Álvarez-Martínez, D. (2021). *Aprendizaje automático en la estimación de pasajeros para una aerolínea de bajo costo*.
- DGAC - PERÚ. (2022). "ANÁLISIS DEL COMPORTAMIENTO DEL TRÁFICO DE PASAJEROS EN EL. DGAC-PERU.
- Enrique, O., & Castillo, M. (2020). *Desarrollo de un modelo para la predicción del precio del cobre empleando herramientas de Machine Learning Tesis para obtener el título profesional de Ingeniera Industrial*.
- Ferrovial. (2022). *Tráfico aéreo: qué es y cómo se controla - Ferrovial*.  
<https://www.ferrovial.com/es/recursos/trafico-aereo/>
- García alonso Sergio. (2021). *MACHINE LEARNING INTEGRATION IN AN SME: CASE STUDY IN DARWINEX*.
- González Wong, E. J. (2019). *CLASIFICACIÓN DE VISUALIZACIONES MOTORAS UTILIZANDO SEÑALES DE UN EEG, BASADO EN ALGORITMOS DE APRENDIZAJE PROFUNDO*.  
<http://repositorio.uach.mx/215/1/TESIS-%20GONZALEZ%20WONG%20EVELYN%20JANETH.PDF>
- Han, J., Kim, S. Y., Lee, J., & Lee, W. H. (2022). *Brain Age Prediction: A Comparison between Machine Learning Models Using Brain Morphometric*

- Data. *Sensors* (Basel, Switzerland), 22(20).  
<https://doi.org/10.3390/s22208077>
- Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-14566-3>
- Kontou, E., Garikapati, V., & Hou, Y. (2020). Reducing ridesourcing empty vehicle travel with future travel demand prediction. *Transportation Research Part C: Emerging Technologies*, 121, 102826.  
<https://doi.org/10.1016/J.TRC.2020.102826>
- Li, Z., Bi, J., & Li, Z. (2017). Passenger Flow Forecasting Research for Airport Terminal Based on SARIMA Time Series Model. *IOP Conference Series: Earth and Environmental Science*, 100(1), 012146.  
<https://doi.org/10.1088/1755-1315/100/1/012146>
- Mago Vistro, D., Rasheed, F., & Gertrude David, L. (2019). The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 8(09). [www.ijstr.org](http://www.ijstr.org)
- Mahesh Batta. (2018). Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Self Flowing Generator View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review. *International Journal of Science and Research*.  
<https://doi.org/10.21275/ART20203995>
- María, I. L., & Quiñones, Z. (2021). *Desarrollo de un modelo predictivo para un sistema de manufactura orientado a la industria 4.0*.
- Moreno Vallejo, P., Bastidas Guacho, G., & Vallejo Sanaguano, M. (2022). Predicción de la demanda de pasajeros mediante un modelo de regresión lineal en R PREDICCIÓN DE LA DEMANDA DE PASAJEROS MEDIANTE UN MODELO DE REGRESIÓN LINEAL EN R. *Junio*, 101–109.
- Nicolás Veiga, Lucía Otero, & Julia Torres. (2020). *Reflexiones sobre el uso de la estadística inferencial en investigación didáctica*. 7(2).  
<https://doi.org/10.29156/INTER.7.2.10>

- Park, Y., Choi, Y., Kim, K., & Yoo, J. K. (2022). Machine learning approach for study on subway passenger flow. *Scientific Reports* 2022 12:1, 12(1), 1–20. <https://doi.org/10.1038/s41598-022-06767-7>
- Puga Reyes. (2020). *UNIVERSIDAD DE GUAYAQUIL FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS CARRERA DE INGENIERIA EN SISTEMAS*. <http://repositorio.ug.edu.ec/bitstream/redug/49617/1/B-CISC-PTG-1832-2020%20Puga%20Reyes%20Eduardo%20Xavier.pdf>
- Ramos-Galarza, C. (2021). Editorial: Diseños de investigación experimental. *CienciAmérica*, 10(1), 1–7. <https://doi.org/10.33210/ca.v10i1.356>
- Rasaizadi, A., Farzin, I., & Hafizi, F. (2022). Machine learning approach versus probabilistic approach to model the departure time of non-mandatory trips. *Physica A: Statistical Mechanics and Its Applications*, 586. <https://doi.org/10.1016/j.physa.2021.126492>
- Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. In *Information (Switzerland)* (Vol. 11, Issue 4). MDPI AG. <https://doi.org/10.3390/info11040193>
- REAL ACADEMIA ESPAÑOLA. (n.d.). *pasajero, pasajera | Definición | Diccionario de la lengua española | RAE - ASALE*. 2022. Retrieved December 14, 2022, from <https://dle.rae.es/pasajero?m=form>
- Rendón Macías, M. E., Villasís Keever, M. Á., & Miranda Novales, M. G. (2016). Descriptive statistics. *Revista Alergia Mexico (Tecamachalco, Puebla, Mexico : 1993)*, 63(4), 397–407. <https://doi.org/10.29262/RAM.V63I4.230>
- Ritha, R., Andre, P., Rui, R., & Carla, M. (2020). *Predicting the Tear Strength of Woven Fabrics via Automated Machine Learning: An Application of the CRISP-DM Methodology*. <https://repositorium.sdum.uminho.pt/bitstream/1822/68602/5/Paper.pdf>
- Rojas, E. M. (2020). *Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo*.

- Rosario, R. (2021). *Comparison of Q-Learning-based reinforcement learning algorithms*.  
<https://riuma.uma.es/xmlui/bitstream/handle/10630/23354/Caponera%20de%20Cobellis%2c%20Romolo%20Rosario%20Memoria.pdf?sequence=1&isAllowed=y>
- Rubio José de Jesús, Hernández Aguilar José Alberto, Ávila-Camacho Francisco Jacob, Stein-Carrillo Juan Manuel, & Meléndez-Ramírez Adolfo. (2016). Sistema sensor para el monitoreo ambiental basado en redes Neuronales. *Ingeniería, Investigación y Tecnología*, 17(2), 211–222.  
<https://doi.org/10.1016/j.riit.2016.06.006>
- Samunderu, E., & Farrugia, M. (2022). Predicting customer purpose of travel in a low-cost travel environment—A Machine Learning Approach. *Machine Learning with Applications*, 9, 100379.  
<https://doi.org/10.1016/j.mlwa.2022.100379>
- Shenzhen Hengyubo. (2022). *Aprende sobre modelado predictivo con cursos online | edX*. Cursos de Modelo Predictivo.  
<https://www.edx.org/es/aprende/modelo-predictivo>
- Universidad de Alcalá. (2022, October 20). *Estilo ISO 690:2010 - Citar y elaborar bibliografía - Bibliogúas at Universidad de Alcalá*. BiblioGúas Biblioteca.  
[https://uah-es.libguides.com/citar\\_elaborar\\_bibliografia/iso](https://uah-es.libguides.com/citar_elaborar_bibliografia/iso)
- Wang, Z., Han, X., Chen, Y., Ye, X., Hu, K., & Yu, D. (2022). Prediction of Willingness to Pay for Airline Seat Selection Based on Improved Ensemble Learning. *Aerospace*, 9(2). <https://doi.org/10.3390/aerospace9020047>
- Xie, X., Leurent, F., & Zhu, Y. (2021). Mining passenger's regional intermodal mobility from smartcard data. *Transportation Research Procedia*, 52, 724–731. <https://doi.org/10.1016/j.trpro.2021.01.087>
- Yang, X., Xue, Q., Ding, M., Wu, J., & Gao, Z. (2021). Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics*, 231. <https://doi.org/10.1016/j.ijpe.2020.107920>

ZAKARYAN, A. (2021). Application of Artificial Intelligence (Neural Networks) in Education. *Main Issues Of Pedagogy And Psychology*, 19(1), 78–87. <https://doi.org/10.24234/miopap.v19i1.395>

Zevallos, F. (2021). *Modelo de predicción para determinar la compra de productos manteniendo un stock adecuado de artículos de Ferretería. Caso aplicado a la Ferretería MAYTA.* [https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/83117/Flores\\_ZMI-SD.pdf?sequence=1&isAllowed=y](https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/83117/Flores_ZMI-SD.pdf?sequence=1&isAllowed=y)

## ANEXOS

### Anexo 1 - Matriz de consistencia de la investigación

Título: Propuesta de un Modelo Predictivo para el Tráfico aéreo de pasajeros en el Perú, 2023

Autor: Pilco Mamani Jhony Antonio

Problema	Objetivo	Hipótesis	Variables
<p>General:</p> <p>¿De qué manera la propuesta de un modelo predictivo influye en el tráfico aéreo de pasajeros en el Perú en el año 2023?</p>	<p>General:</p> <p>Proponer un modelo de predictivo para la toma de decisiones del tráfico aéreo de pasajere en el Perú 2023;</p>	<p>El Desarrollo de un modelo predictivo del tráfico aéreo de pasajeros mejora la toma de decisiones en los aeropuertos del Perú en el año 2023</p>	<p>Independiente:</p> <p>Modelo predictivo</p>
<p>Específicos:</p> <p>1). ¿De qué medida Machine Learning permite predecir el modelo predictivo de tráfico aéreo de pasajeros?</p> <p>2). ¿De qué medida permite predecir la precisión de los modelos de Machine Learning para predecir el tráfico aéreo de pasajeros?</p> <p>3). ¿De qué medida Machine Learning permite determinar el error absoluto del modelo de tráfico aéreo de pasajeros?</p>	<p>Específicos:</p> <ul style="list-style-type: none"> <li>▪ Determinar el tiempo de ejecución de algoritmos de Machine Learning más apropiado para el tráfico aéreo de pasajeros.</li> <li>▪ Determinar la métrica de precisión del modelo predictivo de Machine Learning para permite predecir el mejor modelo del tráfico aéreo de pasajeros.</li> <li>▪ Determinar el error porcentual absoluto medio para establecer el desempeño del modelo predictivo del tráfico aéreo de pasajeros.</li> </ul>	<p>Específico:</p> <ul style="list-style-type: none"> <li>▪ Machine Learning permite predecir el tiempo de ejecución de los algoritmos del tráfico aéreo de pasajeros.</li> <li>▪ Machine Learning permite predecir la precisión del modelo predictivo para predecir el tráfico aéreo de pasajeros.</li> <li>▪ Machine Learning permite predecir el error absoluto del modelo del tráfico aéreo de pasajeros.</li> </ul>	<p>Dependiente:</p> <p>Tráfico aéreo</p>

Metodología

<p>Tipo de investigación:</p> <ul style="list-style-type: none"> <li>• Aplicada</li> </ul>	<p>Población (N):</p> <p><math>N = 980</math></p>	<p>Técnicas de recolección de datos:</p> <ul style="list-style-type: none"> <li>• Observación</li> <li>• Encuesta</li> </ul>	<p>Método de análisis de datos:</p> <ul style="list-style-type: none"> <li>• Estadística descriptiva</li> <li>• Estadística inferencial</li> <li>• Deductivo (enfoque cuantitativo)</li> </ul>
<p>Diseño de investigación:</p> <p>Preexperimental</p>	<p>Muestra (n):</p> <p><math>n = 980</math></p>	<p>Instrumentos de recolección de datos:</p> <ul style="list-style-type: none"> <li>• Ficha de observación</li> <li>• Cuestionario</li> </ul>	<p>Aspectos éticos:</p> <p>Se respetará el derecho a la propiedad intelectual (Originalidad de la investigación - Reporte Turnitin).</p> <p>Se tomará en cuenta el Código de ética de la Universidad César Vallejo.</p> <p>Adicionalmente, se usará para la redacción de la investigación el Sistema de Normas ISO.</p>

## Anexo 2 - Matriz de operacionalización de variables

Variable	Definición Conceptual	Definición Operacional	Dimensión (Sub variable)	Indicador	Métricas de medición
Independiente:  Modelo predictivo	“es un conjunto de estrategias que predicen resultados futuros utilizando el aprendizaje automático, la minería de datos históricos, el big data y el reconocimiento de patrones para la mejora de la toma de decisiones mediante técnicas de análisis de datos” (Shenzhen, 2022).	Se puede medir a través de análisis de los datos como: el procesamiento de los datos usando machine learning, para la toma de decisiones.			
Dependiente:  Tráfico aéreo	“movimiento continuo que se produce en el aire. Necesita una supervisión y gestión periódicas a través de una serie de procesos y un estricto conjunto de directrices”(Ferrovial, 2022)	Se puede medir por el análisis exploratorio para utilizar un modelo predictivo y analizar los datos y predecir el tráfico aéreo para la toma de decisión.	Métricas de precisión	<p style="text-align: center;">Tiempo</p> <hr/> $R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$ <hr/> $MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	<p style="text-align: center;">Ordinal</p> <hr/> <p style="text-align: center;">Razón</p>



### Anexo 3 - Elección de la metodología de desarrollo

*Apellidos y nombres del experto:* Torres Villanueva, Marcelino

*Título profesional y/o Grado académico:* Ingeniero de Sistemas - Maestro.

*Fecha:* 18/11/2022

*Título del proyecto de investigación:* "Propuesta de un Modelo Predictivo para el Tráfico aéreo de pasajeros en el Perú, 2023"

*Autor:* Pilco Mamani, Jhony Antonio

#### Evaluación de la metodología de modelo un predictivo Machine Learning

Mediante el Método de juicio experto, Usted tiene la facultad de calificar las metodologías involucradas, mediante unas series de criterios con puntuaciones especificadas al final de la tabla. Así mismo le exhortamos en la correcta determinación de la metodología para desarrollar la solución propuesta en el presente proyecto de investigación y, también si hubiese algunas sugerencias:

Item	Criterios	Metodologías		
		CRIPS-DM	SEMMA	KDD
1	Compresión del Tema	3	2	1
2	Comprensión de datos	3	1	1
3	Preparación de los datos	3	2	3
4	Modelado	3	1	2
5	Evaluación	2	1	1
Total		14	7	8

La escala a evaluar es de: **1** - Malo, **2** - Regular, **3** – Bueno

Sugerencias:

Firma del experto

#### Criterios de evaluación de las metodologías propuestas

Ítem	Criterio	Descripción
1	Compresión del Tema	Determinar los objetivos del tema de investigación.
2	Comprensión de datos	Es la exploración de los datos.
3	Preparación de los datos	Es el procesamiento y procesamiento de datos.
4	Modelado	Es la construcción del modelo.
5	Evaluación	Es el resultado y revisión del proceso.

Fuente: Elaboración propia

Apellidos y nombres del experto: Guevara Ruíz, Ricardo Manuel

Título profesional y/o Grado académico: Ingeniero Computación y Sistemas - Maestro.

Fecha: 18/11/2022

Título del proyecto de investigación: "Propuesta de un Modelo Predictivo para el Tráfico aéreo de pasajeros en el Perú, 2023"

Autor: Pilco Mamani, Jhony Antonio

### Evaluación de la metodología de modelo un predictivo Machine Learning

Mediante el Método de juicio experto, Usted tiene la facultad de calificar las metodologías involucradas, mediante unas series de criterios con puntuaciones especificadas al final de la tabla. Así mismo le exhortamos en la correcta determinación de la metodología para desarrollar la solución propuesta en el presente proyecto de investigación y, también si hubiese algunas sugerencias:

Item	Criterios	Metodologías		
		CRIPS-DM	SEMMA	KDD
1	Compresión del Tema	3	2	1
2	Comprensión de datos	3	1	1
3	Preparación de los datos	3	2	3
4	Modelado	3	1	2
5	Evaluación	2	1	1
Total		14	7	8

La escala a evaluar es de: **1** - Malo, **2** - Regular, **3** – Bueno

Sugerencias:



Firma del experto

### Criterios de evaluación de las metodologías propuestas

Ítem	Criterio	Descripción
1	Compresión del Tema	Determinar los objetivos del tema de investigación.
2	Comprensión de datos	Es la exploración de los datos.
3	Preparación de los datos	Es el procesamiento y procesamiento de datos.
4	Modelado	Es la construcción del modelo.
5	Evaluación	Es el resultado y revisión del proceso.

Fuente: Elaboración propia

Apellidos y nombres del experto: Agreda Gamboa, Everson David

Título profesional y/o Grado académico: Ingeniero de Sistemas - Doctor.

Fecha: 18/11/2022

Título del proyecto de investigación: "Propuesta de un Modelo Predictivo para el Tráfico aéreo de pasajeros en el Perú, 2023"

Autor: Pilco Mamani, Jhony Antonio

### Evaluación de la metodología de modelo un predictivo Machine Learning

Mediante el Método de juicio experto, Usted tiene la facultad de calificar las metodologías involucradas, mediante unas series de criterios con puntuaciones especificadas al final de la tabla. Así mismo le exhortamos en la correcta determinación de la metodología para desarrollar la solución propuesta en el presente proyecto de investigación y, también si hubiese algunas sugerencias:

Item	Criterios	Metodologías		
		CRIPS-DM	SEMMA	KDD
1	Compresión del Tema	3	2	1
2	Comprensión de datos	3	1	1
3	Preparación de los datos	3	2	3
4	Modelado	3	1	2
4	Evaluación	2	1	1
Total		14	7	8

La escala a evaluar es de: **1** - Malo, **2** - Regular, **3** – Bueno

Sugerencias:

Firma del experto

### Criterios de evaluación de las metodologías propuestas

Ítem	Criterio	Descripción
1	Compresión del Tema	Determinar los objetivos del tema de investigación.
2	Comprensión de datos	Es la exploración de los datos.
3	Preparación de los datos	Es el procesamiento y procesamiento de datos.
4	Modelado	Es la construcción del modelo.
5	Evaluación	Es el resultado y revisión del proceso.

Fuente: Elaboración propia

Anexo 4. Ficha de registro - Instrumento de observación para el algoritmo

Tipo de Prueba	Post Test
Investigador	Jhony Antonio Pilco Mamani
Fecha de inicio	

Algoritmo	Regresión Lineal
-----------	------------------

Matriz de confusión:

N°	Indicador	Medida	Formula	Precisión
1	Precisión	Razón	Modelo	100%
2	Especificidad	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	100%
3	Sensibilidad	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	100%

Otras Métricas

N°	Indicador	Medida	Formula	Precisión
1	F1 SCORE	Razón	$2 \times \frac{Recall * Precision}{Recall + Precision}$	1

Anexo 4A – Cuestionarios para el tráfico aéreo

Nº	Factor	Aspecto	Valor
1	Datos de origen	¿de qué año es el tráfico aéreo?	Dato Numérico
2		¿de qué mes son los datos del tráfico aéreo?	Dato Numérico
3		¿cuál es el nombre de la concesión?	Dato de caracteres
4		¿Cuál es el nombre del aeropuerto?	Dato de Caracteres
5		¿qué tipo de tráfico es?	Dato de Caracteres
6		¿cuántos pasajeros desembarcan?	Dato Numérico
7		¿Cuántos pasajeros embarcan?	Dato Numérico
8		¿Cuál es la latitud del aeropuerto?	Dato Numérico
9		¿Cuál es la longitud del aeropuerto?	Dato Numérico

Anexo 5. Algoritmos de Machine Learning Modelos predictivos comparados para la investigación

TheilSen Regression	Linear Regression	K Neighbors Regressor	AdaBoost Regressor	Decision Regressor	SVM
Utiliza una generalización de la mediana en múltiples dimensiones y es robusto a los valores atípicos multivariantes	Es el enfoque para ajustar un modelo lineal minimizando la suma residual de cuadrados entre el valor observado y el valor predicho por el modelo de regresión por mínimos cuadrados ordinarios.	este algoritmo utiliza la similitud de características para predecir los valores de cualquier nuevo punto de datos, lo que significa que al nuevo punto se le asigna un valor basado en lo mucho que se parece a los puntos del conjunto de entrenamiento.	Se trata de un algoritmo de boosting ensemble en el que los modelos se añaden secuencialmente y los modelos posteriores corrigen las predicciones de los modelos anteriores.	Se trata de un algoritmo de toma de decisiones que utiliza estructura de árbol similar a un diagrama de flujo. Este algoritmo observa características de un objeto que entrenan un modelo en la estructura de un árbol para predecir datos en el futuro y producir una salida	se caracteriza por el uso de kernels, sparsity, control del margen de tolerancia, y el número de vectores de soporte. SVR admite tanto la regresión lineal como la no lineal.

Fuente: (Han et al., 2022)

## Anexo 6. Resultado del modelo Predictivo de Machine Learning

En este anexo, analizamos los resultados de nuestra investigación en función a los modelos predictivos que se entrenaron TheilSen Regressor, Linear Regression, K Neighbors Regressor, AdaBoost Regressor, Decision Tree Regressory Support Vector Regression, los seis algoritmos de aprendizaje para definir el mejor, la validación de hipótesis se realizó utilizando la Regresión Lineal  $R^2$ . Esta función es un proceso estadístico que establece relaciones entre características dependientes e independientes. Así mismo se muestra de la siguiente manera la tabla de los algoritmos de aprendizaje automático utilizados en la investigación.

A continuación de muestra los resultados de cada algoritmo obtenidas a partir de la herramienta de Python y pycaret para modelar cada algoritmo.

- Regresión de TheilSen (tr)

**Tabla 10:** Tabla de Regresión de TheilSen ajuste de parámetro

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	826.0106	3.379528e+06	1838.3492	0.9965	1.4584	6.8750
1	332.7253	2.230771e+05	472.3103	0.9994	1.5471	7.4886
2	1019.7927	7.580526e+06	2753.2754	0.9948	1.3251	6.7652
3	915.4256	7.722030e+06	2778.8541	0.9961	1.1408	3.3369
4	659.7961	2.658442e+06	1630.4729	0.9955	1.3771	5.5942
Mean	750.7501	4.312721e+06	1894.6524	0.9964	1.3697	6.0120
Std	240.0764	2.920089e+06	850.3016	0.0016	0.1370	1.4713

Fuente: Elaboración propia

**Tabla 11:** Tabla de predicción de prueba/rechazo

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
TheilSen Regression	631.605	2.743589e+06	1656.3783	0.9965	1.4429	58791

Fuente: Elaboración propia

Para el entrenamiento del modelo de Regresión de TheilSen se obtuvo una media (*Mean*) 0.9964 y una desviación (*Std*) de 0.0016 de  $R^2$ , para mejor el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión de 0.9964 y una desviación de 0.0016 lo que indica que el modelo es óptimo y en la predicción

de prueba y los resultados de la predicción con los ajustes se tiene un resultado más óptimo de 0.9965 con comparación 0.9964 a anteriores entrenamientos del algoritmo

- Regresión Lineal (lr)

**Table 12:** Tabla regresión lineal ajuste de parámetros

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1111.5283	3.948566e+06	1987.0998	0.9959	1.7924	16.2900
1	615.1142	6.383318e+05	798.9567	0.9983	1.8824	13.5912
2	1110.0563	6.657437e+06	2580.2009	0.9954	1.6063	15.1805
3	1062.0190	6.794740e+06	2606.6721	0.9965	1.6753	9.9843
4	927.1862	3.193239e+06	1786.9637	0.9946	1.7451	11.6599
Mean	965.1808	4.246463e+06	1951.9787	0.9961	1.7403	13.3412
Std	187.4963	2.303178e+06	660.4862	0.0013	0.0950	2.2907

Fuente: Elaboración propia

**Table 13:** Tabla regresión lineal de predicción de prueba/rechazo

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Linear Regression	835.7028	3.063982e+06	1750.4234	0.996	1.8022	14.4061

Fuente: Elaboración propia

En el entrenamiento del modelo de Regresión lineal se obtuvo una media (Mean) 0.9961 y una desviación (Std) de 0.0012 de  $R^2$ , para mejor el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión de la media 0.9961 una desviación de 0.0013 lo que indica que hubo cambio de la desviación en comparación al primero modelo entrenado y en la predicción de prueba con los resultados de la predicción con los ajustes se tiene un resultado más óptimo de 0.996 con comparación 0.9965 a anteriores entrenamientos del algoritmo



- K - vecinos (K-nn)

**Table 14:** Tabla regresión K-nn - ajuste de parámetros

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1009.4457	1.320037e+07	3633.2318	0.9863	0.8760	0.8580
1	183.2132	1.310301e+05	361.9808	0.9997	0.8807	2.0893
2	1007.0515	8.118420e+06	2849.2840	0.9944	0.8373	3.0507
3	1742.0993	4.060588e+07	6372.2744	0.9793	0.5663	0.7178
4	1216.1985	1.253813e+07	3540.9218	0.9787	0.7870	1.9263
Mean	1031.6016	1.491877e+07	3351.5386	0.9877	0.7895	1.7284
Std	501.9441	1.366287e+07	1919.8842	0.0082	0.1166	0.8598

Fuente: Elevación propia

**Table 15:** Tabla regresión K-nn de predicción de prueba/rechazo

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
K Neighbors Regressor	860.3418	5.880623e+06	2424.9996	0.9924	0.9471	3.642

Fuente: Elaboración propia

En el entrenamiento del modelo de K-vecinos(Knn) se obtuvo una media (Mean) 0.9867 y una desviación (Std) de 0.0081 de  $R^2$ , para mejor el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión más optima de la media 0.9877 y una desviación de 0.0082 lo que indica que hubo cambio de la desviación en comparación al primero modelo entrenado y en la predicción de prueba con los resultados de la predicción con los ajustes se obtuvo un resultado más óptimo de 0.9924 con comparación 0.9877 a anteriores entrenamientos del algoritmo

- Regresión AdaBoost (ada)

**Table 16:** Tabla de Regresión AdaBoost ajuste de parámetros

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1690.2969	1.749778e+07	4183.0348	0.9818	1.9820	17.5808
1	1021.5204	2.633037e+06	1622.6634	0.9931	2.3495	28.4356
2	1620.5223	6.847768e+06	2616.8241	0.9953	1.7335	17.6522
3	2447.5334	4.167744e+07	6455.8065	0.9788	1.8745	16.6549
4	1597.5630	1.017823e+07	3190.3346	0.9827	1.9401	18.9509

<b>Mean</b>	1675.4872	1.576685e+07	3613.7327	0.9863	1.9759	19.8549
<b>Std</b>	454.5417	1.383932e+07	1645.5357	0.0066	0.2049	4.3522

Fuente: Elevación propia

**Table 17:** Tabla de Regresión AdaBoost predicción de prueba/rechazo

<b>Model</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>R2</b>	<b>RMSLE</b>	<b>MAPE</b>
AdaBoost Regressor	1825.3342	1.416078e+07	3763.0809	0.9817	2.0052	19.8687

Fuente: Elaboración propia

En el entrenamiento del modelo de Regresión AdaBoost se obtuvo una media (Mean) 0.9868 y una desviación (Std) de 0.0055 de  $R^2$ , para mejorar el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión más optima de la media 0.9963 lo que indica que hubo una reducción y una desviación de 0.0066 lo que indica que hubo cambio de la desviación en comparación al primero modelo entrenado y en la predicción de prueba con los resultados de la predicción con los ajustes se obtuvo un resultado más óptimo de 0.9817 con comparación 0.9863 a anteriores entrenamientos del algoritmo

- Árbol de decisión (dt)

**Table 18:** Tabla Árbol de decisión - ajuste de parámetros

<b>Fold</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>R2</b>	<b>RMSLE</b>	<b>MAPE</b>
<b>0</b>	1101.2154	1.670985e+07	4087.7688	0.9826	0.9017	0.3258
<b>1</b>	362.6930	1.019888e+06	1009.8948	0.9973	0.8772	2.1689
<b>2</b>	883.4697	5.308219e+06	2303.9573	0.9963	0.8837	2.6366
<b>3</b>	1502.5243	2.627915e+07	5126.3193	0.9866	0.5479	0.7081
<b>4</b>	1597.5812	2.091796e+07	4573.6157	0.9645	0.6840	1.5259
<b>Mean</b>	1089.4967	1.404701e+07	3420.3112	0.9855	0.7789	1.4731
<b>Std</b>	447.1174	9.486815e+06	1532.4769	0.0119	0.1401	0.8651

Fuente: Elaboración propia

**Table 19:** Tabla Árbol de decisión predicción de prueba/rechazo

<b>Model</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>	<b>R2</b>	<b>RMSLE</b>	<b>MAPE</b>
Decision Tree Regressor	1022.4528	9.213356e+06	3035.351	0.9881	0.8068	3.4119

Fuente: Elaboración propia

En el entrenamiento del modelo de Árbol de decisión se obtuvo una media (Mean) 0.9806 y una desviación (Std) de 0.0162 de  $R^2$ , para mejorar el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión más óptima de la media 0.9855 lo que indica que hubo una reducción y una desviación de 0.0119 lo que indica que hubo cambio de la desviación en comparación al primer modelo entrenado y en la predicción de prueba con los resultados de la predicción con los ajustes se obtuvo un resultado más óptimo de 0.9881 con comparación 0.9855 a anteriores entrenamientos del algoritmo

- Máquina de vectores de soporte (svm)

**Table 20:** Tabla SVM ajuste de parámetros

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	10469.6792	1.050617e+09	32413.2256	-0.0935	2.5391	23.0415
1	4810.0557	3.963927e+08	19909.6139	-0.0391	2.6924	34.2203
2	14011.2598	1.626119e+09	40325.1612	-0.1234	2.5095	21.0665
3	15320.9369	2.178661e+09	46676.1336	-0.1091	2.6176	18.6276
4	9231.6868	6.599076e+08	25688.6660	-0.1194	2.5702	22.8379
Mean	10768.7237	1.182340e+09	33002.5601	-0.0969	2.5857	23.9587
Std	3720.0624	6.477030e+08	9652.4888	0.0307	0.0642	5.3700

Fuente: Elaboración propia

**Table 21:** Tabla SVM - predicción de prueba/rechazo

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Support Vector Regression	8623.9317	8.336598e+08	28873.1678	-0.0769	2.4885	22.7192

Fuente: Elaboración propia

En el entrenamiento del modelo de Máquina de vectores de soporte se obtuvo una media (Mean) 0.1017 y una desviación (Std) de 0.0306 de  $R^2$ , para mejorar el modelo y obtener los óptimos parámetros más próximos se obtuvo una precisión más óptima de la media 0.0969 lo que indica que hubo una reducción y una desviación de 0.0307 lo que indica que hubo cambio de la desviación en comparación al primer modelo entrenado y en la predicción de prueba con los resultados de la predicción con los ajustes se obtuvo un resultado más óptimo de 0.0769 con comparación 0.0769 a anteriores entrenamientos del algoritmo

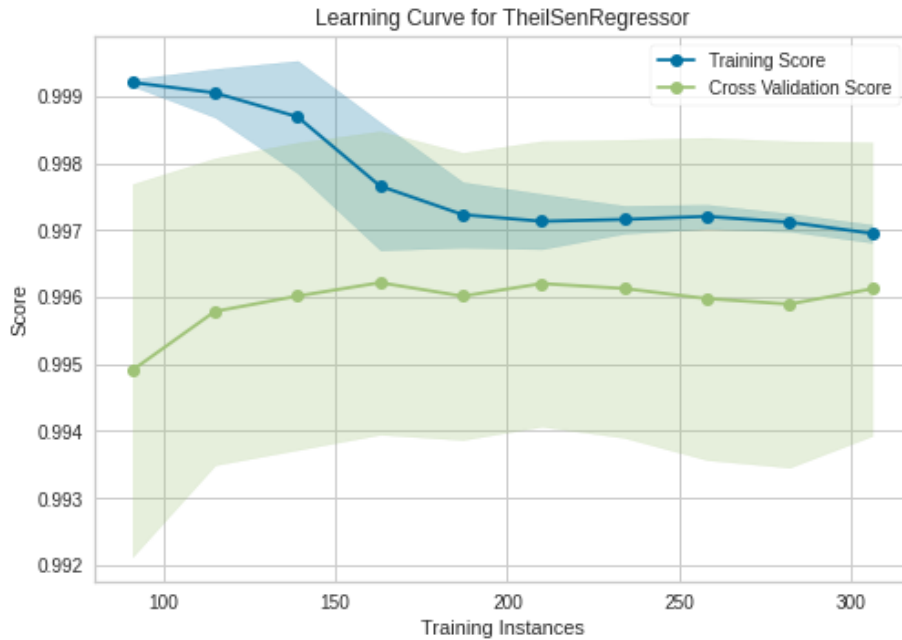
**Tabla 22:** Modelo predictivo de tráfico aéreo

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT(Sec)
tr	TheilSen Regressor	731.1588	4.184067e+06	1735.6893	0.9961	1.3134	5.7634	4.269
Lr	Linear Regression	951.8026	4.304317e+06	1838.5717	0.9950	1.7553	13.0857	0.015
knn	K Neighbors Regressor	1031.8336	1.675634e+07	3069.5992	0.9847	0.8023	1.9057	0.013
ada	AdaBoost Regressor	1782.7913	1.454152e+07	3346.5982	0.9828	2.2110	34.5442	0.035
dt	Decision Tree Regressor	1090.6897	1.758657e+07	3443.5641	0.9826	0.8187	2.5686	0.010
svm	Support Vector Regression	10793.0122	1.184766e+09	30982.7987	-0.1184	2.6296	24.5168	0.017

Fuente: Elaboración propia

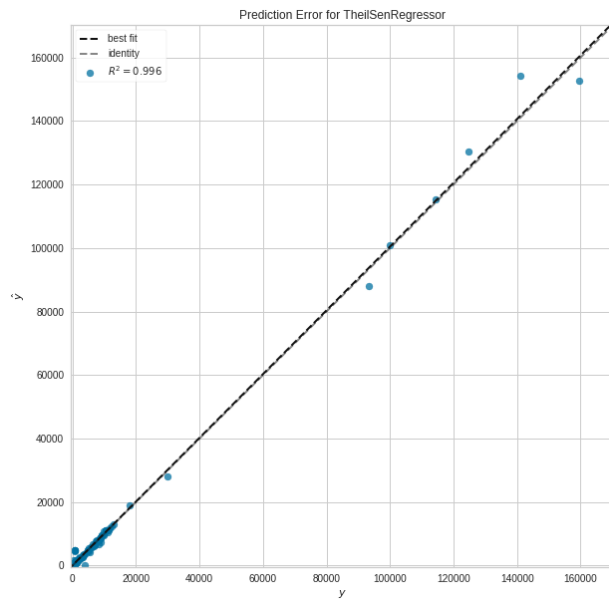
En la siguiente tabla, se puede observar los diferentes modelos entrenados el modelo con mejor resultado de  $R^2$  es TheilSen Regressor con un resultado de 0.9961 fue el que más se aproxima a 1 ya el indica que es un algoritmo óptimo para el modelo predictivo. En comparación con los demás modelos con las que comparo.

Figura 4. Curva de aprendizaje modelo de análisis discriminante lineal



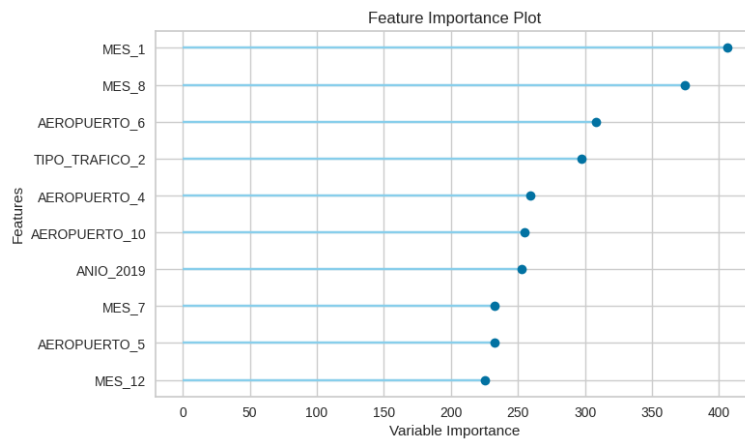
La figura 6 representa la curva de aprendizaje para el análisis discriminante lineal, donde se observó el efecto de observaciones en el rendimiento del modelo. Para nuestro modelo de tráfico de pasajeros observamos que la estabilización de la curva después de 9000 iteraciones muestra un excelente resultado.

Figura 5. Gráfico de curva del error de predicción



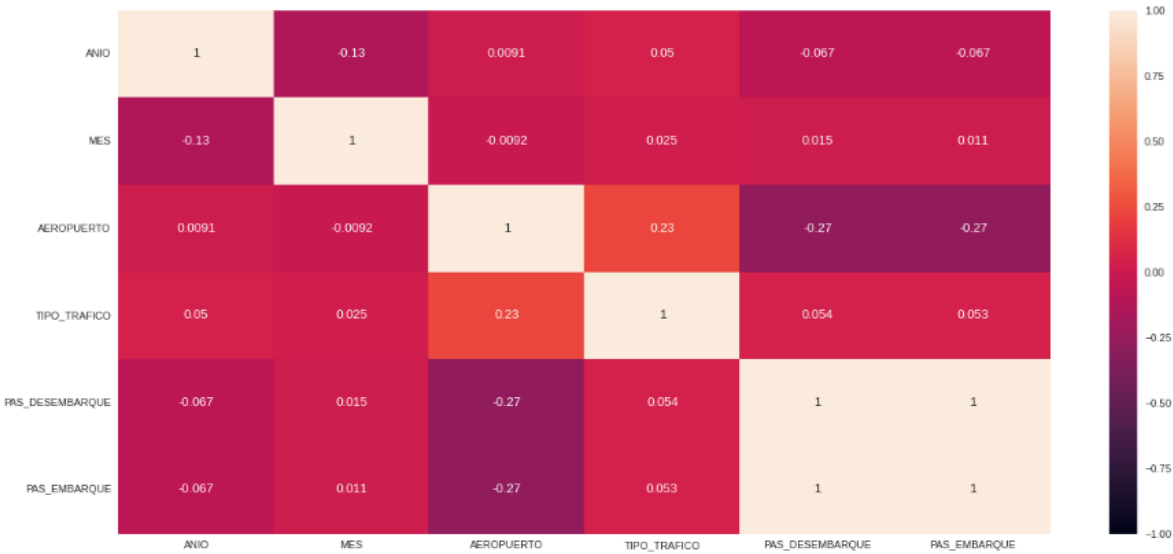
En la figura 2 representa la curva del error de predicción donde se puede observar que el mejor ajuste (best fit) y la identidad(identity) comparten la misma recta, observamos que  $R^2$  tiene un resultado del 0.996, esto hace que el modelo se ajusta bien a los datos en comparación otros modelos para el análisis de tráfico aéreo.

Figura 6. Gráfico modelo de análisis discriminante lineal



En la figura 3 muestra la importancia de las variables categóricas para la construcción del modelo seleccionado. Para esta casa la categoría más notable fue el parámetro de mes de enero (MES\_1), las demás categorías se reflejaron en los 10, mas importantes, con excepción de Pasajeros desembarque y pasajero embarque.

Figura 7. Correlación de variables



En la figura 4 se detalla la correlación de variables del dataset que cuenta con una correlación de la variable tipo tráfico con un valor de 0.23 el valor que se acerca a uno.

Anexo 7 - instrumentos de observación para el algoritmo de Regresión de TheilSen

Tipo de Prueba	Post test
Investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Regresión de TheilSen
-----------	-----------------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	tr
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	0.9965
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	5.8791

Otras Métricas

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE
Regresión Theilsen	631.605	2.743589e+06	1656.3783	0.9965	1.4429	58791

## Anexo 8 - instrumentos de observación para el algoritmo de Regresión Lineal

Tipo de Prueba	Post test
investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Regresión lineal
-----------	------------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	lr
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	0.996
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	14.4061

Otras Métricas

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Regresión lineal	835.7028	3.063982e+06	1750.4234	0.996	1.8022	14.4061



Anexo 9 - instrumentos de observación para el algoritmo de Regresión knn

Tipo de Prueba	Post test
Investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Regresión Knn
-----------	---------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	Knn
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	0.9924
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	3.642

Otras Métricas

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE
Regresión KNN	860.3418	5.880623e+06	2424.9996	0.9924	0.9471	3.642

Anexo 10 - instrumentos de observación para el algoritmo de Regresión de AdaBoost

Tipo de Prueba	Post test
Investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Regresión de AdaBoost
-----------	-----------------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	ada
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	0.9817
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	19.8687

Otras Métricas

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE
Regresión AdaBoost	1825.3342	1.416078e+07	3763.0809	0.9817	2.0052	19.8687

Anexo 11 - instrumentos de observación para el algoritmo de Regresión de TheilSen

Tipo de Prueba	Post test
investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Regresión de árbol de decisión
-----------	--------------------------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	dt
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	0.9881
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	3.4119

Otras Métricas

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Árbol de decisión	1022.4528	9.213356e+06	3035.351	0.9881	0.8068	3.4119

Anexo 12 - instrumentos de observación para el algoritmo de Máquina de Vector de soporte

Tipo de Prueba	Post test
investigador	Jhony Antonio Pilco Mamani

Algoritmo Utilizado:

Algoritmo	Máquina de vector de soporte
-----------	------------------------------

Métricas a evaluar:

Ítem	Indicador	Medida	Formula	Precisión
1	Predicción	Ordinal	Modelo	svm
2	Métrica R2	Razón	$R^2 = 1 - \frac{MSE(Model)}{MSE(baseline)}$	-0.0769
3	MAPE	Razón	$MAPE = \frac{1}{n} \sum_{T=1}^n \left  \frac{A_t - F_t}{A_t} \right $	22.7192

Otras Métricas

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE
SVM	8623.9317	8.336598e+08	28873.1678	-0.0769	2.4885	22.7192

## Anexo 13 - Dataset de tráfico aéreo analizado para el modelo predictivo

202209-PDE-REP-00075.xlsx - Excel

Archivo Inicio Insertar Disposición de página Fórmulas Datos Revisar Vista Ayuda

Calibrí 11 Fuente Alineación Número Estilos Celdas Edición

AN	MES	CONCESION	ENTIDAD PRESTADORA	AEROPUERTO	TIPO TRAFICO	PASAJEROS DESEMBARQU	PASAJEROS EMBARQU
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ATALAYA	NACIONAL	1,062	940
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CHIMBOTE	NACIONAL	0	0
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	INTERNACIONAL	5,636	5,919
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	NACIONAL	132,213	144,514
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	HUÁNUCO	NACIONAL	1,080	1,184
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ILO	NACIONAL	0	0
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JAÉN	NACIONAL	12,446	13,806
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JAUJA	NACIONAL	7,223	7,501
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JUANJUI	NACIONAL	0	0
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	MAZAMARI	NACIONAL	3	5
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	NAZCA	NACIONAL	7,569	7,572
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	RIOJA	NACIONAL	0	0
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	TINGO MARIA	NACIONAL	1,070	1,114
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	TOCACHE	NACIONAL	15	14
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	YURIMAGUAS	NACIONAL	691	783
2019	1	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	sin aginar	NACIONAL	16	20
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ATALAYA	NACIONAL	945	918
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CHIMBOTE	NACIONAL	0	0
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	INTERNACIONAL	4,733	5,148
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	NACIONAL	129,277	124,601
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	HUÁNUCO	NACIONAL	1,458	1,507
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ILO	NACIONAL	538	790
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JAÉN	NACIONAL	11,380	12,184
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JAUJA	NACIONAL	7,997	8,164
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	JUANJUI	NACIONAL	0	0
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	MAZAMARI	NACIONAL	0	8
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	NAZCA	NACIONAL	5,703	5,710
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	RIOJA	NACIONAL	0	11
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ROD. DE MENDOZA	NACIONAL	0	0
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	TINGO MARIA	NACIONAL	1,061	1,079
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	TOCACHE	NACIONAL	55	54
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	YURIMAGUAS	NACIONAL	580	636
2019	2	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	sin aginar	NACIONAL	3	5
2019	3	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	ATALAYA	NACIONAL	1,165	1,009
2019	3	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CHIMBOTE	NACIONAL	0	0
2019	3	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	INTERNACIONAL	5,489	5,405
2019	3	AEROPUERTOS NO CONCESIONADOS	CORPORACIÓN PERUANA DE AEROPUERTOS Y AVIACIÓN COMERCIAL S.A. - CORPAC S.A.	CUSCO	NACIONAL	135,144	135,626

202209-PDE-REP-00075

Listo

Anexo 14. Valores asignados por cada variable del dataset para el procesamiento de datos

<b>MES</b>	
Enero	1
Febrero	2
Marzo	3
Abril	4
Mayo	5
Junio	6
Julio	7
Agosto	8
Septiembre	9
Octubre	10
Noviembre	11
Diciembre	12
<b>AEROPUERTO</b>	
ATALAYA	1
CHIMBOTE	2
CUSCO	3
CUSCO	4
HUÁNUCO	5
ILO	6
JAÉN	7
JAUJA	8
JUANJUI	9
MAZAMARI	10
NAZCA	11
RIOJA	12
TINGO MARIA	13
TOCACHE	14
YURIMAGUAS	15
ROD. DE MENDOZA	16
TOCACHE	17
SAPOSOA	18
<b>TIPO TRAFICO</b>	
Nacional	1
Internacional	2

### Anexo 15. Transformación de los datos del Dataste del tráfico aéreo de pasajeros

	A	B	C	D	E	F	G	H	I	J	K
1	ANIO	MES	AEROPUERTO	TIPO_TRAFICO	PAS_DESEMBARQUE	PAS_EMBARQUE					
2	2019	1	2	2	1,062	940					
3	2019	1	4	1	5,636	5,919					
4	2019	1	4	2	132,213	144,514					
5	2019	1	5	2	1,080	1,184					
6	2019	1	7	2	12,446	13,806					
7	2019	1	8	2	7,223	7,501					
8	2019	1	10	2	3	5					
9	2019	1	12	2	7,569	7,572					
10	2019	1	16	2	1,070	1,114					
11	2019	1	17	2	15	14					
12	2019	1	18	2	691	783					
13	2019	2	2	2	945	918					
14	2019	2	4	1	4,733	5,148					
15	2019	2	4	2	129,277	124,601					
16	2019	2	5	2	1,458	1,507					
17	2019	2	6	2	538	790					
18	2019	2	7	2	11,380	12,184					
19	2019	2	8	2	7,997	8,164					
20	2019	2	10	2	0	8					
21	2019	2	12	2	5,703	5,710					

## Anexo 16. Código fuente Machine Learning del modelo predictivo del tráfico aéreo de pasajero

```
from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

# Demora 1'54"
!pip install pycaret

#Se necesita mas adelante en el setup del PyCaret
!pip install numba

Looking in indexes: https://pypi.org/simple, https://us-
python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: numba in /usr/local/lib/python3.8/dist-
packages (0.54.1)
Requirement already satisfied: numpy<1.21,>=1.17 in
/usr/local/lib/python3.8/dist-packages (from numba) (1.19.5)
Requirement already satisfied: llvmlite<0.38,>=0.37.0rc1 in
/usr/local/lib/python3.8/dist-packages (from numba) (0.37.0)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.8/dist-packages (from numba) (57.4.0)

import pandas as pd

from pycaret.utils import enable_colab
enable_colab()

Colab mode enabled.

df = pd.read_excel("/content/drive/MyDrive/Colab
Notebooks/Tesis/Data_set_Trafico_Aereo.xlsx")
df.head()

<IPython.core.display.HTML object>

   ANIO  MES  AEROPUERTO  TIPO_TRAFICO  PAS_DESEMBARQUE  PAS_EMBARQUE
0  2019    1             2              2             1062             940
1  2019    1             4              1             5636             5919
2  2019    1             4              2            132213            144514
3  2019    1             5              2             1080             1184
4  2019    1             7              2            12446            13806

# Variable dependiente
df.columns

<IPython.core.display.HTML object>

Index(['ANIO', 'MES', 'AEROPUERTO', 'TIPO_TRAFICO', 'PAS_DESEMBARQUE',
       'PAS_EMBARQUE'],
      dtype='object')

# Valores faltantes
df.isnull().sum()
```



```
<IPython.core.display.HTML object>
```

```
ANIO          0
MES           0
AEROPUERTO    0
TIPO_TRAFICO  0
PAS_DESEMBARQUE 0
PAS_EMBARQUE  0
dtype: int64
```

```
# informacion de las columnas, y vemos que son numericas
df.info()
```

```
<IPython.core.display.HTML object>
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 542 entries, 0 to 541
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  ---                ---
0   ANIO                  542 non-null    int64
1   MES                   542 non-null    int64
2   AEROPUERTO           542 non-null    int64
3   TIPO_TRAFICO         542 non-null    int64
4   PAS_DESEMBARQUE     542 non-null    int64
5   PAS_EMBARQUE         542 non-null    int64
dtypes: int64(6)
memory usage: 25.5 KB
```

```
# Observemos los tipos de columnas de nuestro dataframe df
# Deseo obtener solo las columnas numericas
ListaVarNumericas = []
for col in df.columns:
    if (df[col].dtype == 'int64') or (df[col].dtype == 'float64') :
        ListaVarNumericas.append(col)
#
ListaVarNumericas = ListaVarNumericas[:]
```

```
<IPython.core.display.HTML object>
```

```
# Asignamos
X = df[ListaVarNumericas]
```

```
<IPython.core.display.HTML object>
```

```
# Verifiquemos la existencia de valores faltantes
X.isnull().sum()
```

```
<IPython.core.display.HTML object>
```

```
ANIO          0
MES           0
AEROPUERTO    0
```

```

TIPO_TRAFICO      0
PAS_DESEMBARQUE  0
PAS_EMBARQUE      0
dtype: int64

#Tamaño del dataset, filas y columnas
dataset=X
dataset.shape

<IPython.core.display.HTML object>

(542, 6)

data = dataset.sample(frac=0.9, random_state=786)
data_unseen = dataset.drop(data.index)

data.reset_index(drop=True, inplace=True)
data_unseen.reset_index(drop=True, inplace=True)

print('Datos para el modelamiento: ' + str(data.shape))
print('Datos no vistos para predicciones:' + str(data_unseen.shape))

<IPython.core.display.HTML object>

Datos para el modelamiento: (488, 6)
Datos no vistos para predicciones:(54, 6)

!pip install NumPy

from pycaret.regression import *
exp_reg101 = setup(data = data, target = 'PAS_EMBARQUE',
session_id=123)

```

### Comparing All Models

```

best = compare_models(exclude = ['ransac'])

print(best)

<IPython.core.display.HTML object>

HuberRegressor(alpha=0.0001, epsilon=1.35, fit_intercept=True,
max_iter=100,
               tol=1e-05, warm_start=False)

models()

!pip install scikit-optimize

plot_model(best,plot='error') ###queda
plot_model(best, plot = 'feature')

```

```
import matplotlib.pyplot as plt
import seaborn as sns

<IPython.core.display.HTML object>

plt.rcParams["figure.figsize"]=(20,10)

sns.heatmap(df.corr(), vmin = -1, vmax = 1, annot = True)
```

### se compararon 6 algoritmos para este modelo

los algoritmos que se evaluaron fueron

- **tr** => TheilSen Regressor
- **Lr** => linear regression
- **KNN** => K Neighbors Regressor
- **ada** => AdaBoost Regressor
- **dt** => Decision Tree Regressor
- **svm** => Support Vector Regression

```
## 'lr','svm','knn','dt','ada', 'tr'
bests = compare_models(include = ['lr','svm','knn','dt','ada', 'tr'])

plot_model(bests, plot= 'learning')
```

### Lr => linear regression

algoritmo de regresion lineal se creo un modelo para evaluarlo

```
lr = create_model('lr', fold = 5)
tuned_lr = tune_model(lr, fold = 5)
```

```
plot_model(tuned_lr)
```

```
plot_model(lr, plot='error')
plot_model(tuned_lr, plot='feature')
plot_model(tuned_lr, plot= 'learning')
predict_model(tuned_lr)
```

### tr => TheilSen Regressor

```
## Creación del modelo TheilSen Regressor
tr = create_model('tr')
```

```
## se tunea el modelo TheilSen Regressor para que ajustar los
hiperparametros del modelo utilizado:
tuned_tr = tune_model(tr, fold = 5)

## Predecimos el modelo final de TheilSen Regressor:
predict_model(tuned_tr)

## Grafico de learning curva de TheilSen Regressor:
plot_model(tuned_tr, plot= 'learning')

## grafico Prediction Error TheilSen Regressor:
plot_model(tuned_tr, plot='error')

## grafico de importancia de la característica TheilSen Regressor:
plot_model(tuned_tr, plot='feature')
```

### **KNN = Algoritmo Knn**

```
## creacion de modelo de algoritmo de Knn
knn = create_model('knn', fold = 5)

## se tunea el modelo KNN para que ajustar los hiperparametros del
modelo utilizado:
```

```
tuned_knn = tune_model(knn, fold = 5)
predict_model(tuned_knn)
plot_model(tuned_knn, plot='error')
plot_model(tuned_knn, plot='learning')
```

### **ada => Algoritmo de Regesion AdaBoost**

```
ada = create_model('ada', fold = 5)
tuned_ada = tune_model(ada, fold = 5)
predict_model(tuned_ada)
plot_model(tuned_ada, plot='error')
plot_model(tuned_ada, plot='feature')
```

### **dt => Algoritmo Arbol de desicion**

```
dt = create_model('dt', fold = 5)
tuned_dt = tune_model(dt, fold = 5)
predict_model(tuned_dt)
plot_model(tuned_dt, plot='error')
plot_model(tuned_dt, plot='feature')
```

### SVM => Algoritmo Máquina de vector de soporte

```
svm = create_model('svm', fold = 5)
tuned_svm = tune_model(svm, fold = 5)
predict_model(tuned_svm)
plot_model(tuned_svm, plot='error')
plot_model(tuned_svm, plot='learning')
```



**UNIVERSIDAD CÉSAR VALLEJO**

**FACULTAD DE INGENIERÍA Y ARQUITECTURA  
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS**

### **Declaratoria de Autenticidad del Asesor**

Yo, AGREDA GAMBOA EVERSON DAVID, docente de la FACULTAD DE INGENIERÍA Y ARQUITECTURA de la escuela profesional de INGENIERÍA DE SISTEMAS de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA ATE, asesor de Tesis titulada: "Propuesta de un Modelo predictivo para el Tráfico aéreo de pasajeros en el Perú, 2023", cuyo autor es PILCO MAMANI JHONY ANTONIO, constato que la investigación tiene un índice de similitud de 14.00%, verificable en el reporte de originalidad del programa Turnitin, el cual ha sido realizado sin filtros, ni exclusiones.

He revisado dicho reporte y concluyo que cada una de las coincidencias detectadas no constituyen plagio. A mi leal saber y entender la Tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

En tal sentido, asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

LIMA, 22 de Febrero del 2023

<b>Apellidos y Nombres del Asesor:</b>	<b>Firma</b>
AGREDA GAMBOA EVERSON DAVID <b>DNI:</b> 18161457 <b>ORCID:</b> 0000-0003-1252-9692	Firmado electrónicamente por: AGREDA el 27-03- 2023 08:57:54

Código documento Trilce: TRI - 0534646