Universidad Del Valle de Guatemala
Inteligencia Artificial CC3045
Sección 10
Ciclo I 2022
24 de mayo del 2022

Proyecto Final
Análisis de transacciones bancarias y su legalidad

Marco Ramírez #19588
Pablo Coutiño #18817

INTRODUCCIÓN **3**

VARIABLES **4**

RESULTADOS CON ARBOL DE DECISION **6**
- Precisión del modelo 6
- Tiempo del modelo 6
- Árbol generado 7

RESULTADOS CON RANDOM FOREST **7**
- Precisión del modelo 7
- Tiempo del modelo 7
- Árbol generado 8

RESULTADOS CON KMEANS **8**

RESULTADOS CON GAUSSIAN MIXTURE MODELS **9**

RESULTADOS **9**

DISCUSIÓN **10**

CONCLUSIONES **10**

REFERENCIAS **11**

INTRODUCCIÓN

El objetivo de este proyecto es analizar las transacciones bancarias y su legalidad, esto mediante la base de datos proporcionada por Kaggle, la cual cuenta con 6,362,620 registros, donde el 99.8% de los datos son no fraudulentos (en este contexto no es tan sorprendente que exista esta disparidad porque no es una actividad que en la que la mayoría de la población incurra por las implicaciones éticas y legales), a través de estos datos buscaremos responder la siguiente pregunta,¿Cuál es el mejor modelo de clasificación que permite predecir de mejor manera una transacción?
Para poder responder la pregunta principal es necesario crear una comparativa entre los diferentes modelos, además de contar con la variable respuesta para poder medir la precisión (accuracy) y tiempo de las clasificaciones. Los modelos que pondremos a prueba serán los siguientes:

- Árboles de decisiones: este es un algoritmo usado mayormente en machine learning, el cual nos permite la construcción de modelos predictivos para el Big Data basados en su clasificación según ciertas características o propiedades, o en la regresión mediante la relación entre distintas variables para predecir el valor de otra. (Unir, 2021).
  La estructura de este árbol es la siguiente:
  - Nodos internos: Estos representan cada una de las características o propiedades a considerar para tomar una decisión.
  - Ramas: Representan la decisión en función de una determinada condición.
  - Nodos finales: Representa el resultado de la decisión.

  Se ha decidido usar este algoritmo debido a su alta tasa de precisión además de generar resultados fáciles de comprender.

- Random forest: Este algoritmo se deriva de los árboles de decisiones, donde ambos son algoritmos de aprendizaje supervisado, con la diferencia de que este se construye a partir de varios árboles, esté establece el resultado en función de las predicciones de los árboles, tomando el promedio o la media de la salida de varios árboles. Por ello random forest se considera más preciso ya que genera mucho más procesos, con la desventaja que requiere de mayor computación. (Mbaabu,2020)

- K-Means: Es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática. El algoritmo consta de tres pasos:
  - Inicialización: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiendo aleatoriamente.
  - Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.
  - Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.(S,2022)

- Gaussian Mixture Models: es un modelo probabilístico que supone que todos los puntos de datos se generan a partir de una mezcla de un número finito de distribuciones gaussianas con parámetros desconocidos, este implementa el algoritmo de maximización de expectativas para ajustar modelos mixtos de Gauss. (Scikit)

- Naive Bayes: es un algoritmo de aprendizaje supervisado, que se basa en el teorema de Bayes y se utiliza para resolver problemas de clasificación. Se utiliza principalmente en la clasificación de texto que incluye un conjunto de datos de entrenamiento de alta dimensión. (Brownlee, 2020)

Cabe mencionar que para la creación de los modelos se requiere preprocesamiento de los datos para que los sets de entrenamiento y test tuvieran representación de transacciones fraudulentas. Se balancearon los conjuntos de entrenamiento y evaluación para que la proporción de registros fraudulentos y legítimos fuera 1:1 y así poder generar un modelo capaz de identificar bien ambos casos con el conjunto de evaluación .

VARIABLES

Step: indica la hora en la que se realizó la transacción desde el inicio de la captura de datos. min: 0; max: 744

type: variable categórica, describe el tipo de transacción que se realizó, puede ser transferencia, débito, crédito , pago o retiro.

amount: variable numérica que indica el monto por el cual se hizo la transacción. Promedio : 179862 ; Min : 0 ; Max: 92,445,517

nameOrig: variable categórica, la cual indica el cliente que inició la transacción.

OldbalanceOrg : variable numérica , la cual indica el balance con el cual inicia la cuenta origen al iniciar la transacción. Min: 0 ; Max: 59585040 ; Promedio: 833883

newbalanceOrg: variable numérica, la cual indica el balance con el cual finaliza la cuenta de origen al concluir la transacción. Min: 0; Max: 49585040 ; Promedio: 855114

nameDest: Variable categórica. Indica quien es el cliente que recibe la transacción.

OldBalanceDest: Variable numérica. Destinatario del saldo inicial antes de la transacción. Min: 0 ; Max: 356015889 ; Promedio: 1100702

newBalanceDest: Variable numérica. Nuevo balance del saldo destino después de la transacción. Min: 0 ; Max: 356179279 ; Promedio: 1224996

isFraud: variable categoría, indica con un 1 sí la transacción es fraudulenta y un 0 sí es legítima.

isFlaggedFraud: variable categórica; se marca con 1 cuando una transacción intenta procesar un movimiento mayor a 200,000.

RESULTADOS CON NAIVE BAYES

Precisión del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Precisión	0.7202	0.7133	0.71333	0.71072	0.717	0.7149

Tiempo del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Tiempo (s)	29.33	14.7238	27.8717	25.2557	18.942	23.22

RESULTADOS CON ARBOL DE DECISION

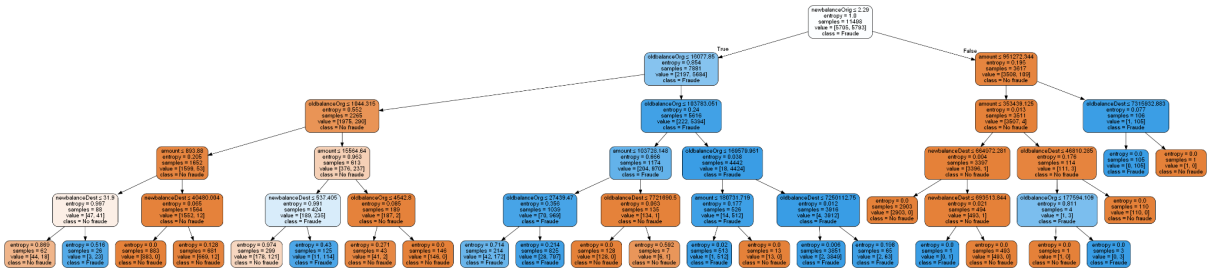
Precisión del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Precisión	0.976663961038961	0.977881493506493 6	0.977272727272727 3	0.981128246753246 7	0.977272727272727 3	0.978043831

Tiempo del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Tiempo (s)	9.647608041763306	9.43937611579895	10.42638826370239 3	10.53209137916565	10.10816502571106	10.03072577

Árbol generado



RESULTADOS CON RANDOM FOREST

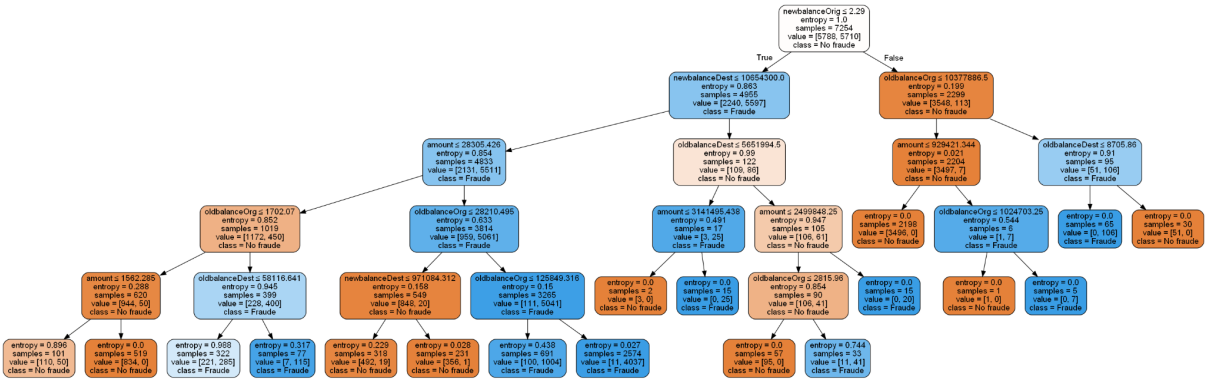
Precisión del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Precisión	0.974634740259740 3	0.969561688311688 3	0.967532467532467 6	0.972808441558441 6	0.971185064935065	0.971144481

Tiempo del modelo

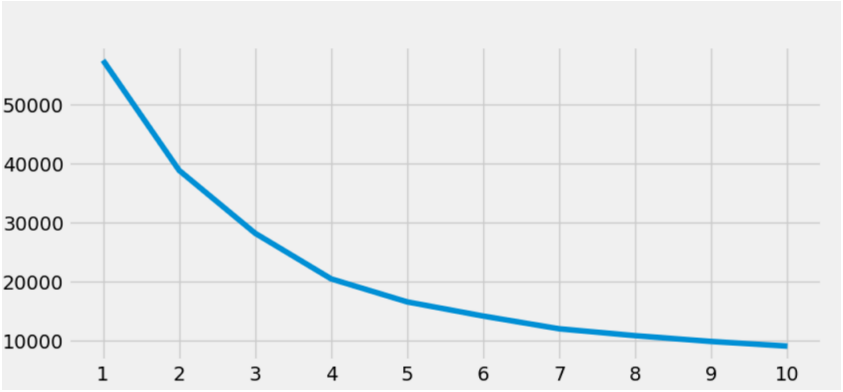
	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Tiempo (s)	11.87778258323669 4	11.39265084266662 6	11.03845500946045	10.80793142318725 6	10.82054734230041 5	11.18747344

Árbol generado



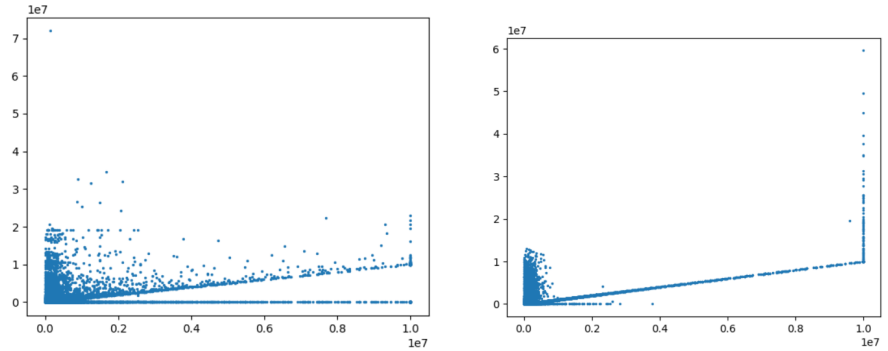
RESULTADOS CON KMEANS

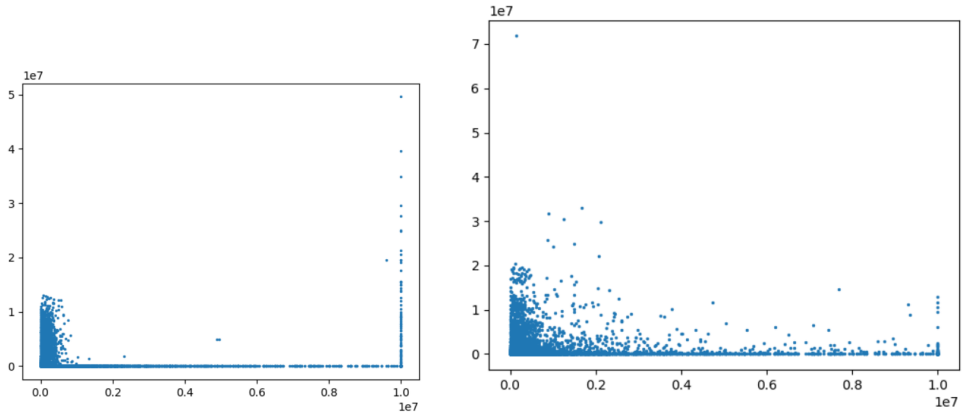
Selección de número de cluster por metodo del codo:



Estadístico de hopkins de : 0.0021213528889209277

Se seleccionaron 4 clusters.





Tiempo del modelo

	Corrida 1	Corrida 2	Corrida 3	Corrida 4	Corrida 5	Corrida Promedio
Tiempo (s)	19.68	15.27	16.61	14.96	17.68	16.84

RESULTADOS

Algoritmo	Precisión promedio	Tiempo promedio
Árboles de decisión	0.978043831	10.03072577
Random Forest	0.971144481	11.18747344
Naive bayes	0.7149	23.22
KMEANS	NA	NA
Mixture Models	NA	NA

VIDEOS EXPLICATIVOS

Random Forest y Decision tree: https://youtu.be/OI6LTmSnNxU

DISCUSIÓN

El estudio y predicción de este fenómeno tuvo su grado de complejidad porque para realizar modelos que hagan buenas predicciones se necesita un buen conjunto de entrenamiento, esto generalmente significa que las categorías en las que se clasificaron los datos deben estar bien representados.
En este caso, las actividades fraudulentas son ocurrencias raras por las implicaciones éticas y legales que tienen, en este dataset la proporción de actividades legítimas a fraudulentas es de 100:1.
Esta disparidad se tomó en cuenta para la segmentación de los registros en el conjunto de entrenamiento y evaluación para que se tuviesen cantidades similares de registros legítimos e ilegítimos en los conjuntos de entrenamiento y de evaluación.
Tras la realización de toma de datos de ambos árboles, se demostró que el árbol más preciso fue árboles de decisión y este también fue el más rápido, sin embargo, la teoría dice que random forest debería ser más preciso, pero mas tardado, ya que que consta de muchas árboles de decisión combinados con el fin de obtener una mejor precisión, es por ello que este requiere de mayor computación. El cual mediante los resultados se demostró lo contrario.

CONCLUSIONES

- Fue clave tener presente el concepto de garbage-in-garbage-out para determinar cómo abordar el problema de la desproporción en los datos y pre-procesarlos para que los conjuntos de entrenamiento y evaluación estuvieran balanceados y tener modelos que generan buenas predicciones.

- En aprendizaje supervisado es posible hacer ajustes a los conjuntos de entrenamiento y evaluación porque las etiquetas de las agrupaciones son conocidas.

- A pesar que random forest tarda más en realizar la predicción con el fin de una mejor precisión, no siempre será la mejor precisión comparada con otros algoritmos.

REFERENCIAS

- Brownlee, J. (2020, August 15). Naive Bayes for Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/naive-bayes-for-machine-learning/
- Javatpoint. (n.d.). Naive Bayes Classifier in Machine Learning - Javatpoint. Www.Javatpoint.Com. https://www.javatpoint.com/machine-learning-naive-bayes-classifier
- MANCHANDA, C. H. I. T. W. A. N. (2022, March 1). Fraudulent Transactions Data. Kaggle. https://www.kaggle.com/datasets/58ea43a05b203eca4de1d23aaed6c819a7625691eaaa033775e31783d70847b6?resource=download
- Mbaabu, O. (2020, December 11). Introduction to Random Forest in Machine Learning. Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/
- McGonagle, J. (n.d.). Gaussian Mixture Model | Brilliant Math & Science Wiki. Brilliant. https://brilliant.org/wiki/gaussian-mixture-model/
- S. (2022, February 25). K-means Clustering Algorithm: Applications, Types, and How Does It Work? Simplilearn.Com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm#:%7E:text=K%2DMeans%20clustering%20is%20an,objects%20belonging%20to%20another%20cluster.
- scikit. (n.d.). 2.1. Gaussian mixture models. Scikit-Learn. https://scikit-learn.org/stable/modules/mixture.html#:%7E:text=A%20Gaussian%20mixture%20model%20is,Gaussian%20distributions%20with%20unknown%20parameters.
- Unir, V. (2021, October 19). Árboles de decisión: en qué consisten y aplicación en Big Data. UNIR. https://www.unir.net/ingenieria/revista/arboles-de-decision/
- Universidad de Oviedo. (n.d.). El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. Unioviedo. https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html