**CLASIFICACIÓN DE TRANSFERENCIAS AL EXTERIOR DE CLIENTES DE ENTIDADES FINANCIERAS EN LA BALANZA DE PAGOS DE BOLIVIA**

Las entidades bancarias reportan transacciones individuales al exterior de sus clientes clasificándolas en nueve categorías relacionadas a la Balanza de Pagos. Debido al volumen de transacciones y la necesidad de su apropiación en las estadísticas externas se hace necesario la automatización de su clasificación mediante técnicas de procesamiento de Lenguaje Natural. Se identificó el Modelo SVM sobre ITFIDF como el mas recomendable para su utilización para generar una aplicación en producción.

1. **Introducción**

La Balanza de pagos es el registro de las transacciones de una economía con no residentes en un determinado periodo de tiempo (FMI, 2009). La publicación de nuevas estadísticas de la balanza de pagos (BoP) divulgadas por el Banco Central de Bolivia (BCB) son importantes para los inversores, los responsables de la formulación de políticas y los economistas académicos con el objeto de analizar y pronosticar el desempeño económico de un país.

Dichas cifras deben estar disponibles de manera oportuna cumpliendo buenas prácticas estadísticas que implican la disminución del error del recurso humano con el consecuente impacto en cuentas de ajuste como Errores y Omisiones.

En particular, las transferencias al exterior del sector privado (familias) se considera de forma agrupada en la (BoP). Sin embargo, la información generada por cada banco sirve para apropiar cada transacción en cuentas externas (Comercio de Bienes, Servicios, Ingreso Primario y Ingreso Secundario, Inversión de cartera y Otra inversión)

Se buscar clasificar transacciones de transferencias al exterior de clientes de bancos comerciales utilizando textos libres adjuntos a cada operación bancaria para una correcta compilación de estadísticas de Balanza de Pagos. El proceso actual es manual con intentos de implementación de soluciones en Excel mediante búsqueda de palabras clave. Si bien el cliente de la EIF expresa el motivo de la transferencia, el registro en el sistema de información la realiza el funcionario bancario con el consiguiente error humano.

1. **Dataset**

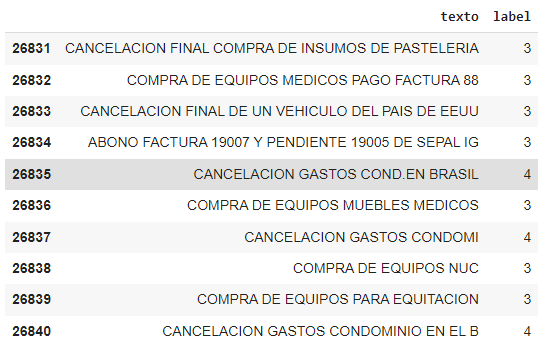
Se trabajó con 26.000 instancias elegidas aleatoriamente de transacciones de frecuencia diaria desde junio de 2021 a marzo de 2022. Las categorías se muestran en el siguiente cuadro y reflejan el lineamiento estadístico para la imputación en cifras del sector externo y sus relaciones con otros sistemas estadísticos publicado por el Banco Central de Bolivia.

**Cuadro 1. Clases del Dataset**



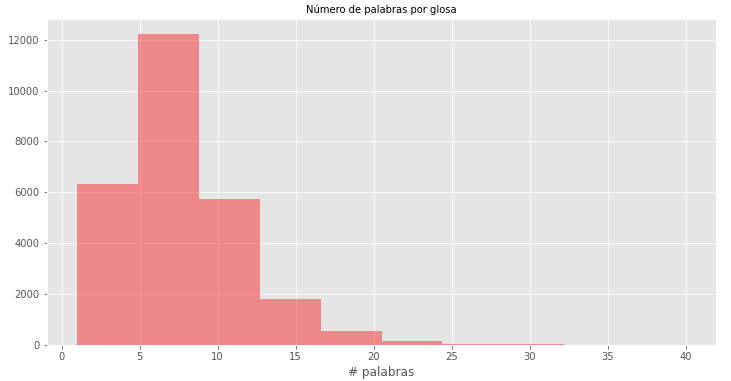
Por confidencialidad, el dataset sólo contiene el texto o glosa que el cliente reporta a la entidad financiera y la clasificación realizada por el analista de balanza cambiaria y mercado cambiario del BCB. El dataset original también contiene datos de la entidad financiera y los montos que por su importancia podrían ser sujetos de futuras investigaciones toda vez que existen indicios de relaciones entre los tipos de transacciones con el tipo de entidad financiera y los montos.

**Cuadro 2. Últimos datos del Dataset**



En el siguiente gráfico se evidencia que la mayoría de los textos de las transacciones tienen menos de 15 palabras.

**Gráfico 2. Número de palabras**



Fuente: BCB

Utilizando la técnica de visualización de datos Word Cloud, a continuación, se exponen las palabras con mayor frecuencia. Destacan palabras como “pago”, “cuenta” y “compra” que son muy comunes en el vocabulario bancario y coloquial en menor grado.

**Gráfico 3. Nube de palabras del texto**



Mediante librerías de ScikitLearn y nltk y librerías en español se tokenizaron los textos, se eliminaron palabras irrelevantes al análisis (stopwords).

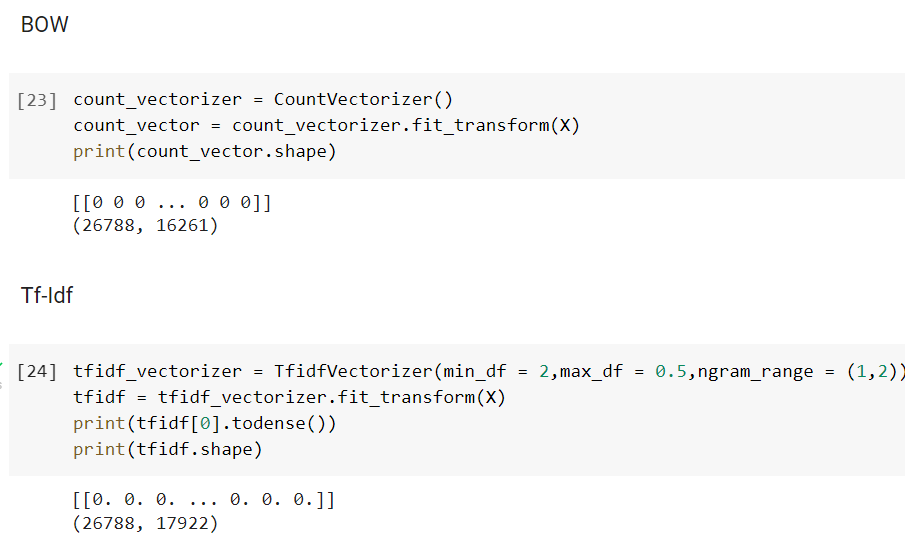
**Cuadro 4. Limpieza de texto**

Mediante la herramienta *CountVectorizer* proporcionada por la biblioteca *scikit-learn* en Python se realizaron transformaciones del texto en una matriz sobre la base de la frecuencia (recuento) de cada palabra que aparece en todo el texto. Esto es útil cuando tenemos varios textos de este tipo y deseamos convertir cada palabra de cada texto en vectores para usar en análisis de texto posteriores y aplicando los clasificadores correspondientes.

El problema con la anterior técnica es que algunas palabras como “pago” aparecerán muchas veces y sus recuentos grandes no serán muy significativos en los vectores codificados. Una alternativa es calcular las frecuencias de las palabras, y el método más popular es el llamado TF-IDF. Este es un acrónimo que significa Frecuencia de Término – Frecuencia Inversa de Documento que son los componentes de las puntuaciones resultantes asignadas a cada palabra.

En el presente proyecto se utilizarán ambas herramientas y ambas matrices sobre las cuales se aplicarán cuatro clasificadores

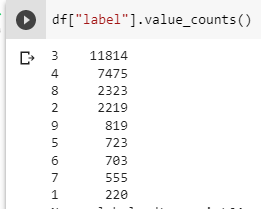
**Cuadro 5. Vectorización del texto**



1. **Exploración de datos**

La categoría 3 “Bienes” contiene gran parte de las transacciones seguida de la categoría “Servicios” e “Inversión en el exterior”. Claramente se trata de un dataset no balanceado que podría acarrear problemas en el momento de escoger las instancias para entrenamiento y prueba.

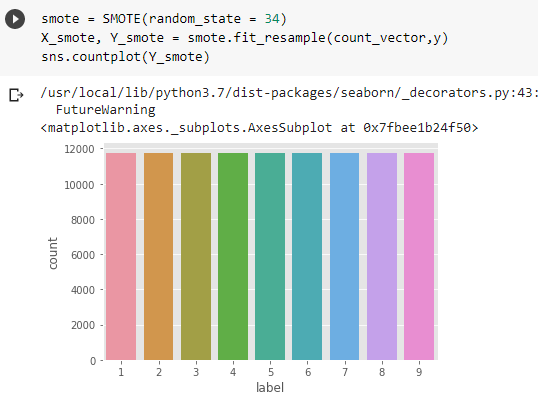
**Cuadro 6. Frecuencia de categorías**



Para subsanar el problema de datos no balanceados se utilizó la herramienta SMOTE que es una técnica estadística de sobremuestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada. El componente funciona cuando genera nuevas instancias a partir de casos minoritarios existentes que se proporcionan como entrada. Esta implementación de SMOTE no cambia el número de casos de mayoría.

A continuación, se muestra que la mencionada herramienta balanceó las instancias para su posterior uso.

**Gráfico 7. Aplicación de SMOTE**

******

Fuente: BCB

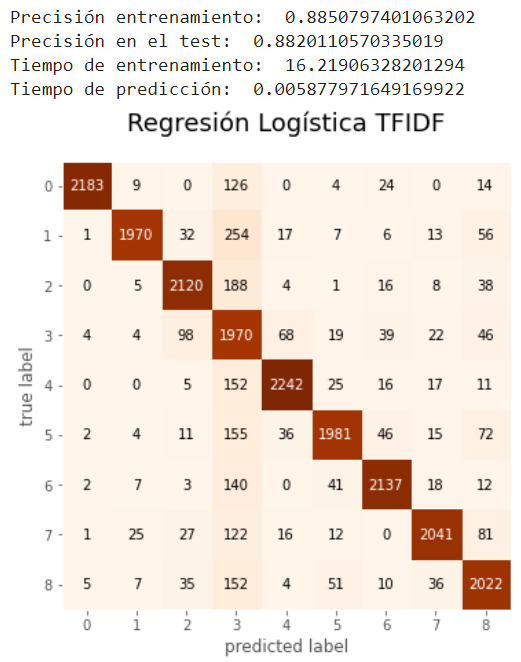
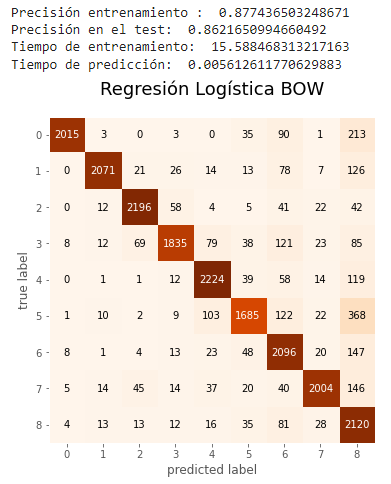
Clasificadores

Regresión logística

El siguiente cuadro muestra los resultados del proceso de entrenamiento sobre las matrices mencionadas líneas arriba. El valor 'C' y también especificar cómo debe manejar el problema de clasificación de clases múltiples (uno contra todos o multinomial). Debido a la categorización de los textos basado en el criterio de diferentes analistas, se determinó un valor C=0,1.

Se puede apreciar la alta precisión en el grupo de entrenamiento y de prueba. El tiempo de entrenamiento es considerable tomando en cuenta que el objetivo de este proyecto es recomendar el mejor clasificador basado en criterios de tiempo y exactitud. Destaca que las categorías con mayores aciertos son “2. Bienes” y “4. Remesas”.

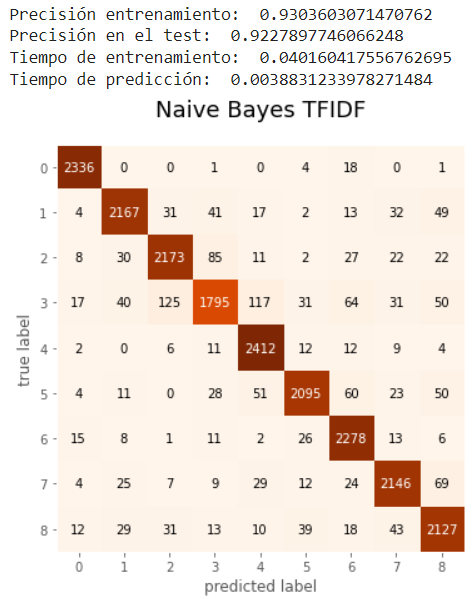
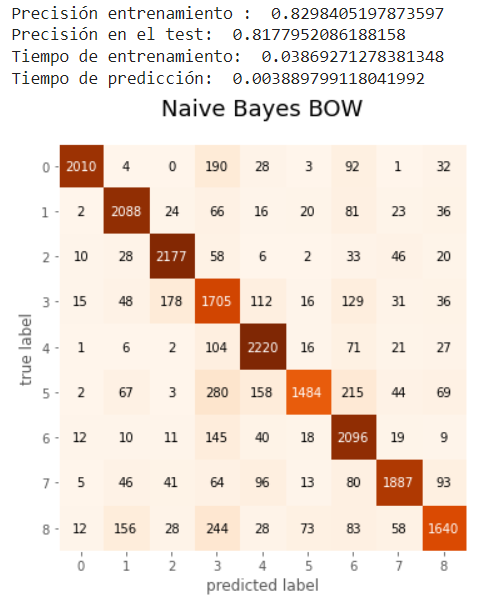
**Gráfico 8. Matriz de confusión, Regresión Logística**



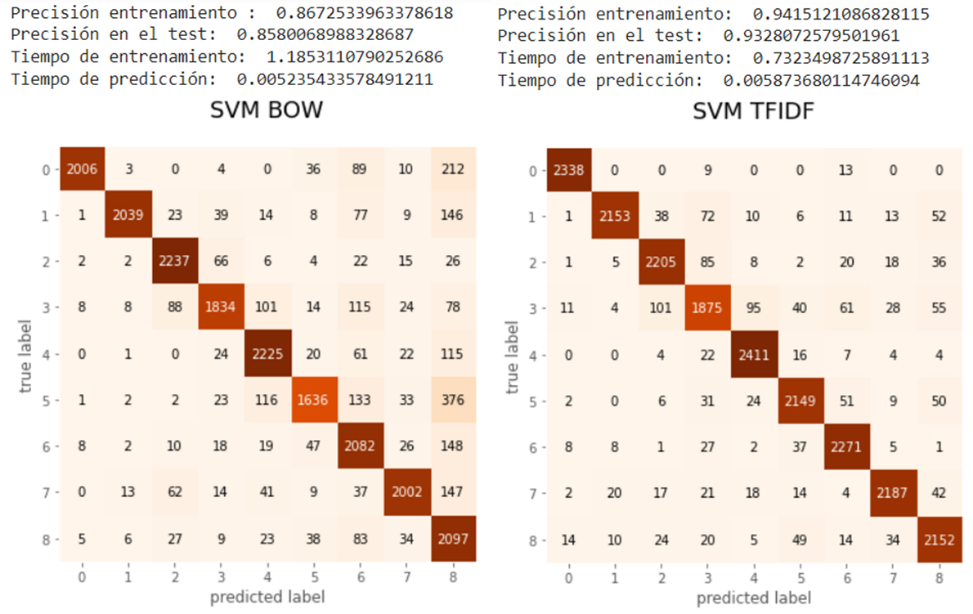
Naive Bayes

Este clasificador pertenece a la familia de clasificadores que se basan en el popular teorema de probabilidad de Bayes, son conocidos por crear modelos simples pero de buen rendimiento, especialmente en nuestro caso, en los campos de clasificación de documentos.

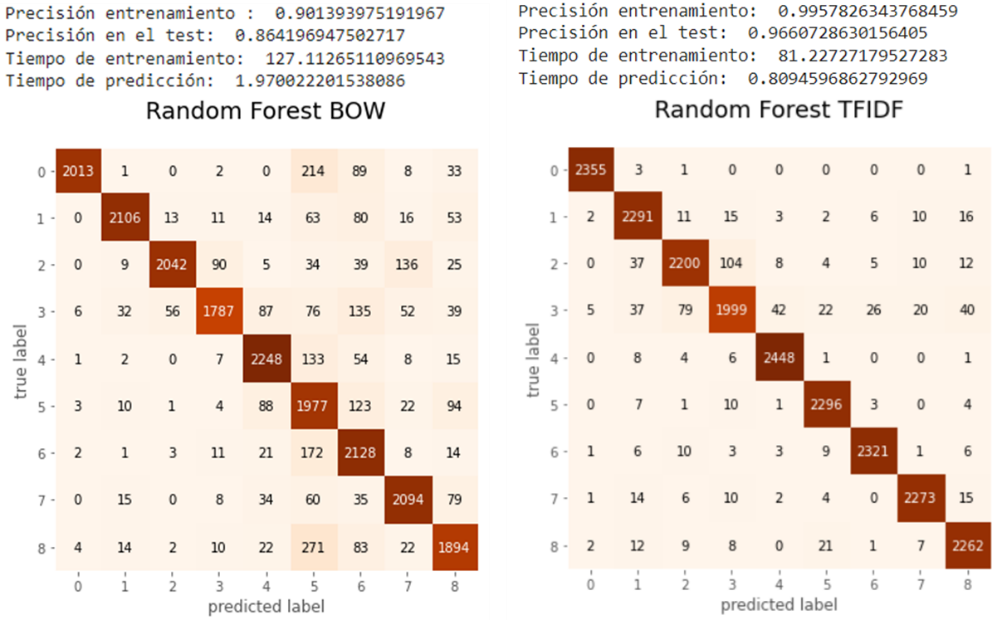
En las siguientes gráficas se aprecia que, a diferencia de los resultados con la aplicación de Regresión Logística, la precisión es menor con el costo en el tiempo.



sdf



sdsdfssd

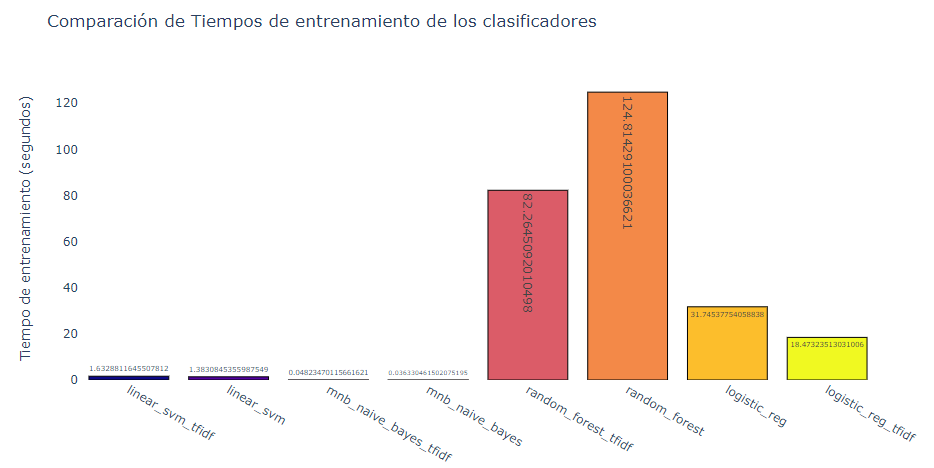


* + 1. **Evaluación**

Es evidente que existe ivalente a $us351millones.

**Gráfico 9. Tiempos de entrenamiento**

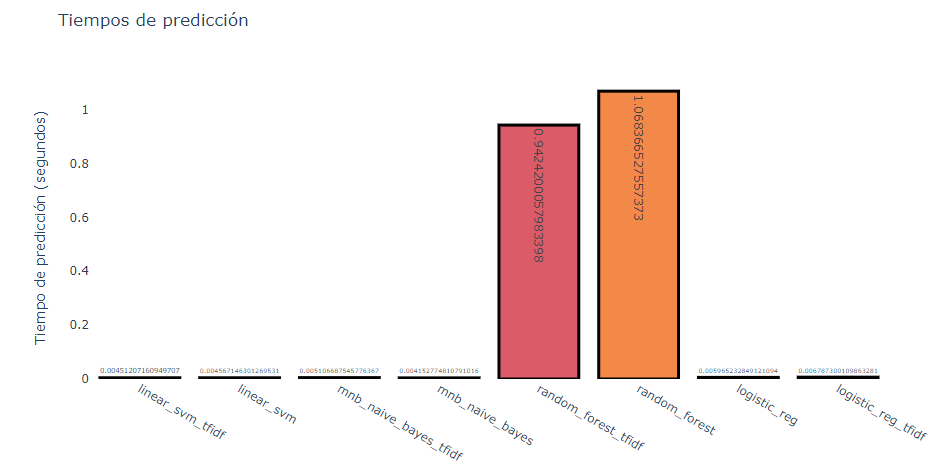
(En segundos)



volumen de operaciones financieras producto de venta de bienes, pago por se

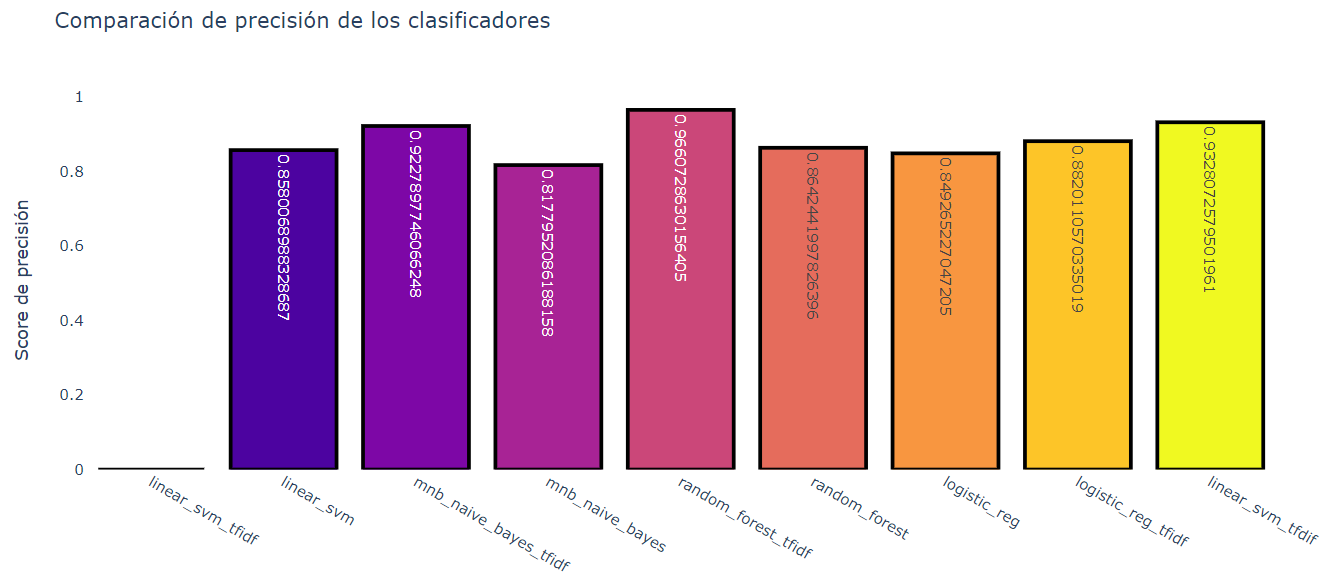
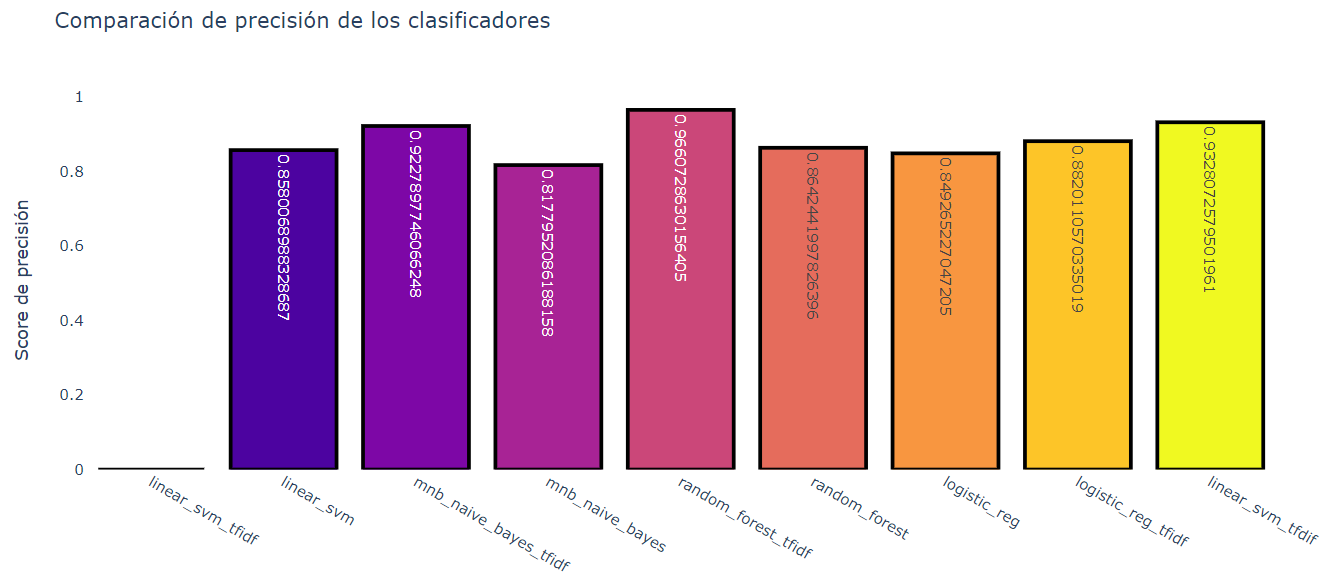
**Gráfico 9. Tiempos de predicción**

(En segundos)



volumen de operaciones financieras producto de venta de bienes, pago por servicios financieros, constitución de activos, etc., la cuenta financiera podría

**Gráfico 9. Precisión**

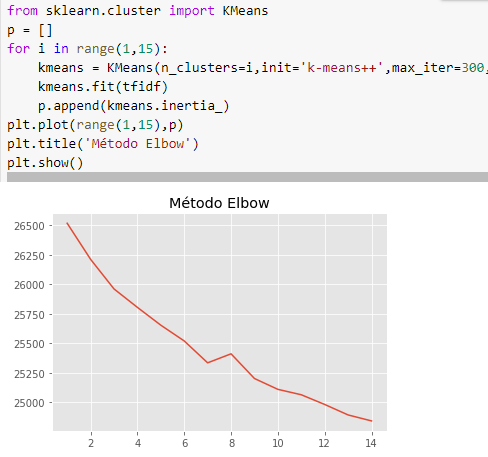


* 1. **Discrepancias en cuenta financiera**

Debido al xtranjera. Se hace notar que las posiciones iniciales y finales en la cuenta de Moneda y

**Cuadro 3.** **Posición de Inversión Internacional: Actual**

(En millones de $us)



1. **Conclusiones**

Se ha logrado implementar 8 modelos de clasificación de textos siendo RandomForest sobre TFIDF el que presenta mayor precisión.

Para la puesta en producción se utilizará el modelo SVM sobre TFIDF por su bajo tiempo de entrenamiento y predicción, y su alta precisión.

1. **Referencia bibliográfica**

Cheong Tang, T. (2013). *New Perspectives on the ‘Net Errors & Omissions’ in Balance of Payment.* Monash University.

Cobham, A., Garcia-Bernardo, J., & Harari, M. (2021). *Vulnerability and exposure to illicit financial flows risk in Latin America.* Buckinghamshire, United Kingdom.: Tax Justice Network Limited.

FMI. (2009). *Manual de Balanza de Pagos y Posición de Inversión Internacional.* FMI.

Global Financial Integrity. (2018). *Trade-Related Illicit Financial Flows.* Global Financial Integrity.

Kristinsson, T. (2016). *Net Errors and Omissions in Balance of Payments Statistics.* NOVA Information Managemente School.

Sarialioglu Hayali, A. (2015). *The Error And Omissions Issue: A Comparison Between Turkey And India.* Bolumu, Trabzon-TURKEY: Technical University Department of Economics.