Implementación de nuevas tecnologías como el machine learning y el análisis de datos para el mercado bancario

Santiago Cárdenas Jiménez
Diego Alejandro Casasbuenas Deaza

Escuela de Ingeniería Julio Garavito, Bogotá, Colombia
santiago.cardenas@mail.escuelaing.edu.co
diego.casasbuenas@mail.escuelaing.edu.co

1 Introducción

Los alcances de la inteligencia artificial hoy por hoy llegado a todas las áreas del mercado que por medio de los algoritmos machine learning que, realizando tareas de clasificación, predicción y agrupamiento, podemos encontrar patrones, tendencias y hasta comportamientos del mercado hacia las ventas, preferencias, segmentación, desarrollo de producto, demanda, y demás [1], [2]. Para este tema, la minería de datos que se encarga del manejo de estas técnicas y algunas otras relacionadas ha desarrollado, implementado y evaluado los algoritmos de tal manera que se usen dentro del mercado, para usos de este artículo, se desarrollara el proceso de minería de datos con una base de datos de un banco portugués, el cual lanzó una campaña de llamadas para la suscripción de depósitos a largo plazo a causa de una disminución de ingresos, para esto se desarrollaron algoritmos de clasificación supervisada, para determinar que perfiles si se suscribirían al producto y no se suscribirían teniendo en cuenta las ventajas de la minería de datos para el desarrollo de este problema, además se desarrolló un algoritmo clustering no supervisado para determinar quiénes si se suscribirían. Dentro de este desarrollo también se incluyó el análisis exploratorio de datos con base en gráficos [3]–[5]. La base de datos contiene información de duración de llamadas, productos abiertos con el banco, estado civil, edad, último contacto, educación, ocupación, frecuencia y deuda, con base a esta información de cada cliente se clasificó y se agrupó.

Palabras clave: Algoritmos, machine learning, bank marketing, classification, clustering.

2 Trabajo relacionado

Como se relacionó en la introducción, podemos ver que al problema a desarrollar es la disminución de ingresos del banco, por lo tanto se propuso la solución de aplicar modelos de predicción para tener una clasificación de personas que se suscribirían al depósito a largo plazo, teniendo en cuenta las ventajas que tiene el manejo de datos, la minería y las nuevas opciones para mejorar el marketing de un banco como se evidencia en la literatura evaluada en el banco UK, para esto se utilizó un algoritmo de clasificación dentro de los cuales se seleccionaron y evaluaron Naive bayes, KNN clasifier, Support Vector Machine [1]–[3]

Además de la clasificación, también se hizo un clustering no supervisado con el método K prototipe para agrupan las características que definen que una persona se suscribiría al depósito a largo plazo. Las ventajas del machine learning es que se puede utilizar en cualquier área, no solo la bancaria, para el caso de la literatura que se buscó para usos de este artículo, tenemos el uso de KNN para predecir posibles problemas durante construcciones, teniendo en cuenta anteriores problemas en anteriores construcciones de los estados unidos, con el fin de optimizar el beneficio evitando los problemas [6]. Además de esto también tenemos un modelo de clasificación de patologías en las cuerdas vocales usando Naive bayes, teniendo en cuenta el tono de voz, se analiza y predice si el paciente tiene alta probabilidad de contraer una enfermedad vocal, también se tienen factores en cuenta como la ocupación debido a que algunas usan las cuerdas vocales más que en otros como en el call center [7].

Por último, se tuvo en cuenta en la literatura un modelo de predicción de incendios forestales teniendo en cuenta la metodología support vector machine (SVM), en este algoritmo se evalúa la probabilidad de que un terreno sea propenso a un incendio forestal teniendo en cuenta características meteorológicas como la humedad, la temperatura, la erosión, y demás[8].

Adicionalmente, se realizó un análisis exploratorio de datos con gráficos para sacar conclusiones adicionales de los datos utilizando gráficos de barras, gráficos de tortas, mapas de calor, mapas de contorno, y demás gráficos para encontrar patrones [5].

El problema en cuestión tiene que ver con la predicción de que una persona se suscriba a un servicio bancario con el fin de incrementar los ingresos del banco, sin embargo, esto no impide que se utilicen las mismas técnicas, al contrario, se encontró su uso y se aplicó al dataset con la información bancaria de cada cliente.

3 Definición del problema y algoritmo

3.1 Problema

En este proyecto se está trabajando con un banco portugués, el cual presenta una disminución en los ingresos, el banco está buscando conocer la razón del problema y saber que acciones tomar. Después de realizar una investigación encontraron que la causa raíz de su problema era que sus clientes no estaban invirtiendo en depósitos a largo plazo en la cantidad que se requiere. Habiendo encontrado esto al banco le gustaría identificar a los clientes que tienen actualmente de forma que se pueda identificar cuales presentan una mayor posibilidad de suscribirse a un depósito de largo plazo y así enfocar las próximas campañas de marketing en dichos clientes.

Los datos que se presentan en el dataset se basan en campañas de marketing del banco las cuales se hicieron en llamadas telefónicas y en ciertas ocasiones presenciales.

3.2 Algoritmos

Los algoritmos utilizados para clasificar a los clientes como aptos a suscribirse o no fueron ‘KNN neighbors’, ‘Naive Bayes’ y ‘Support Vector Machine’, y se evaluó el que tuviera mejores métricas haciendo distintas iteraciones y aplicando técnicas para mejorar los modelos.

El ‘KNN neighbors’ es un modelo predictivo que recorre los datos y calcula la distancia entre ellos, una vez la calcula, este busca los k vecinos más cercanos, dependiendo de la distancia y entre el grupo de los 5 más cercanos, el algoritmo escoge la clase más común dentro del grupo y la asigna al dato buscado, esta es una herramienta que no solo se usa para la clasificación, sino que también se usa para imputación. Adicionalmente se aplicó la técnica ‘cross-validation’ para mejorar este modelo el cual itera con parámetros de manera aleatoria y selecciona la mejor combinación [1], [6].

Además, se usó el algoritmo ‘Naive Bayes’, el cual tiene varias variantes, dependiendo del tipo de dato que se esté utilizando se utiliza una distribución de probabilidad diferente, entre estas variantes se encuentran: Gaussiana, Bernoulli, y Multinomial. La Gaussiana utiliza una distribución de gauss especial para datos numéricos, la Bernoulli para datos boléanos y la multinomial para datos discretos. Este algoritmo se basa en la probabilidad condicional y el teorema de bayes para identificar la clase a la que pertenece con mayor probabilidad de que ocurra [1], [7].

El último algoritmo de clasificación que se usó fue el ‘SVM’ el cual se basa en trazar vectores para separar los datos de una clase y los datos de otra, obteniendo finalmente el un hiperplano que maximice el margen entre cada clase y de esta manera se clasifique dependiendo de su ubicación con respecto al hiperplano. Este algoritmo al igual que los anteriores dos, tiene la técnica ‘Cross-Validation’ [1], [8].

Por último, en la aplicación de algoritmos de clustering se hizo uso de ‘K-prototypes’, este algoritmo tiene como objetivo agrupar el conjunto de datos en un número K de grupos minimizando la función de costo. Una de las principales características de ‘K-prototypes’ es que a diferencia de otros algoritmos de clustering tales como K-means, este se puede aplicar a datasets que contengan variables categóricas mixtas.[9]

4 Evaluación experimental

4.1 Data

El data set utilizado está relacionado con el mundo bancario, específicamente los datos provienen de un banco de Portugal, este data set como se mencionó anteriormente muestra los datos de los clientes del banco, y la información recopilada sobre los clientes a través de las campañas de marketing realizadas. El data set contiene 21 columnas y 32950 filas; cada uno de los conjuntos de datos están ordenados por fecha. El conjunto de datos tiene como objetivo, lograr la clasificación de los clientes, prediciendo si estos se suscribirán en un depósito a largo plazo.

Las características del data-set se presentan a continuación en la Tabla 1:

Caracteristicas	Tipo de dato	Descripción
age	numérica	Edad de la persona
Job	Categórica-nominal	Tipo de trabajo que realiza la persona('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
Marital	Categórica-nominal	Estado civil ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
Education	Categórica-nominal	Nivel de educación ('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
Default	Categórica-nominal	¿El cliente tiene un crédito en mora? ('no','yes','unknown')
Housing	Categórica-nominal	¿El cliente tiene un préstamo de vivienda? ('no','yes','unknown')
Loan	Categórica-nominal	¿El cliente tiene un préstamo personal? ('no','yes','unknown')
Contact	Categórica-nominal	Tipo de comunicación('cellular','telephone')
Month	Categórica-nominal	Ultimo contacto con el cliente en un mes especifico de cadad año('jan', 'feb', 'mar', …, 'nov', 'dec')
Day of week	Categórica-nominal	Dia del ultimo contacto decon el cliente('mon','tue','wed','thu','fri')
Duration	numérica	Duración del último contacto con el clienteen segundos(e.g., if duration=0 then y='no')
Campaign	numérica	Numero de contactos realizados durante esta campaña y para este cliente
Previous	numérica	Numero de contactos realizados antes de esta campaña y para este cliente
poutcome	Categórica-nominal	Resultado de la campaña de marketing anterior('failure','nonexistent','success')

Tabla 1, características de la data set

Como se muestra en la tabla 1 las características del dataset más importantes se centran en la interacción que tuvo la campaña de marketing sobre el cliente, tales como el mes, el día, la duración del contacto con el cliente y la cantidad de veces que se tuvo un contacto con este, además de esto una de las características más relevantes del dataset es el “poutcome” el cual muestra si la campaña de marketing tuvo éxito al aplicar a él diverso cliente

4.2 Metodología

4.2.1 Preprocesamiento

En la fase de preprocesamiento se identificaron los datos que contenía el data set, posteriormente al encontrar datos faltantes se realizó la imputación de estos, después de esto con el objetivo de obtener un data set en las condiciones que plantea el proyecto requiere, se realizó la eliminación del 10% de los datos. A partir de lo anterior se volvió a realizar la imputación sobre los datos nulos, obteniendo un data set equilibrado con base a la moda de los datos utilizando la librería scikit learn, por último, para obtener una mejor calidad de los datos, se aplico un método de normalización sobre los datos numéricos. Además, se discretizaron las variables categóricas para aplicar los modelos de clasificación con la librería Pandas.

Cabe resaltar que para este dataset se tuvo en cuenta un 70% de datos para entrenar y 30% para testear, por lo cual para dividir estos datos aleatoriamente utilizamos la librería panda.

4.2.2. Análisis exploratorio de datos

Para el análisis exploratorio de datos utilizamos las librerías matplotlib. pyplot, seaborn y plotly con el fin de desarrollar gráficos dinámicos que nos ayuden a evidenciar varios hallazgos.

4.2.3 Clasificadores

Supply Vector machine:

Para aplicar el siguiente clasificador se revisó la librería scikit learn, la cual tiene todo tipo de información acerca de los parámetros y uso de este clasificador con Python, este clasificador tiene código llamado SVC y parámetros como coeficiente de regularización, kernel, grado, coeficiente gamma, probabilidad, y peso de la clase. Para lograr mejorar el modelo de clasificación, se utilizó la técnica cross-validation para iterar varias veces con una malla de parámetros aleatorios escogiendo el de mejor métrica, para esto se utilizó la librería scikit learn, la cual tiene la función 'gridSearchCV ()' la cual tiene como parámetros un diccionario con los parámetros que va a iterar la función [10], [11].

Naive bayes:

Para este modelo de clasificación se utilizó la librería scikit learn la cual tiene todo tipo de documentación en cuanto a los tipos de bayes como lo es Gaussian, Bernoulli y multinomial, cada función en Python tiene parámetros de 'priors' que prioriza un atributo y 'var_smoothing' que adiciona la varianza más larga para calcular la estabilidad. De estas funciones se utilizó 'GaussianNB', 'BernoulliNB' y 'MultinomialNB' [10], [12].

KNN neighbors:

El modelo de KNN se utilizó de la librería scikit learn obteniendo los parámetros y optimizándolo con la función gridSearchCV (), dentro de la cual iteró la mejor cantidad de vecinos, el algoritmo utilizado y la métrica de evaluación [10], [13].

4.2.3 Clustering

K-Protoypes:

El modelo aplicado con K-Prototypes se realizó a partir de la librería de kmodes, esta librería de clustering se basa en el metodo de asignación de puntos más cercanos, a partir de los cuales se encuentran los clústers que representan de mejor manera los datos, por otro lado, la personalización es una de las principales características de kmodes, ya que permite escoger los centroides de los clusters con base a diferentes parámetros, como la media o la moda. [9]

Evaluación:

Para la evaluación se utilizó la librería scikit learn importando métricas y utilizando matriz de resultados y matriz de confusión para evidenciar que resultados tiene cada modelo utilizado [10].

4.3 Resultados

Análisis exploratorio de datos

Comenzando con el análisis exploratorio de datos, se evidenció que la población evaluada era mayoritariamente del 64.3% casadas, 25.5% solteros y 10.2% divorciados como evidenciamos en la 'Ilustración 1'.

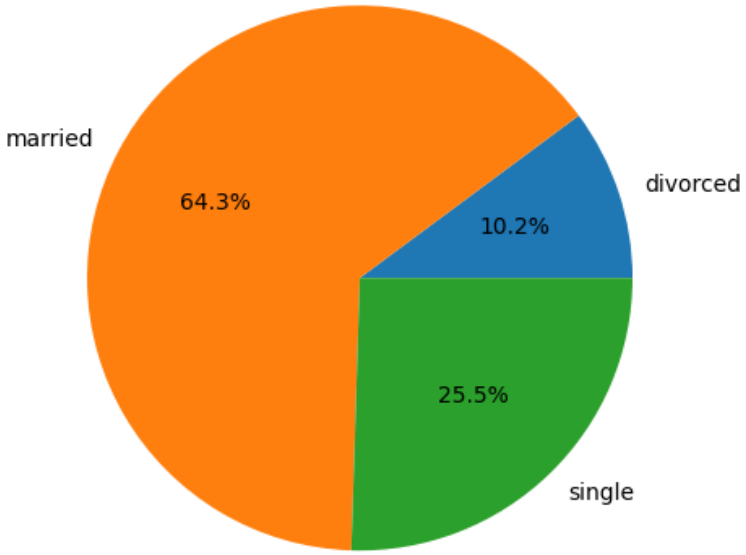

Ilustración 1. Gráfico de torta

Haciendo un análisis de la edad con respecto a la duración de las llamadas, podemos evidenciar que las personas que tienen créditos en el banco suelen tener más duración que los que no tienen crédito, además podemos evidenciar que en cuanto las personas tienen más edad, la duración de las llamadas aumenta. Estos hallazgos se pueden evidenciar en la 'Ilustración 2' e 'Ilustración 3'.

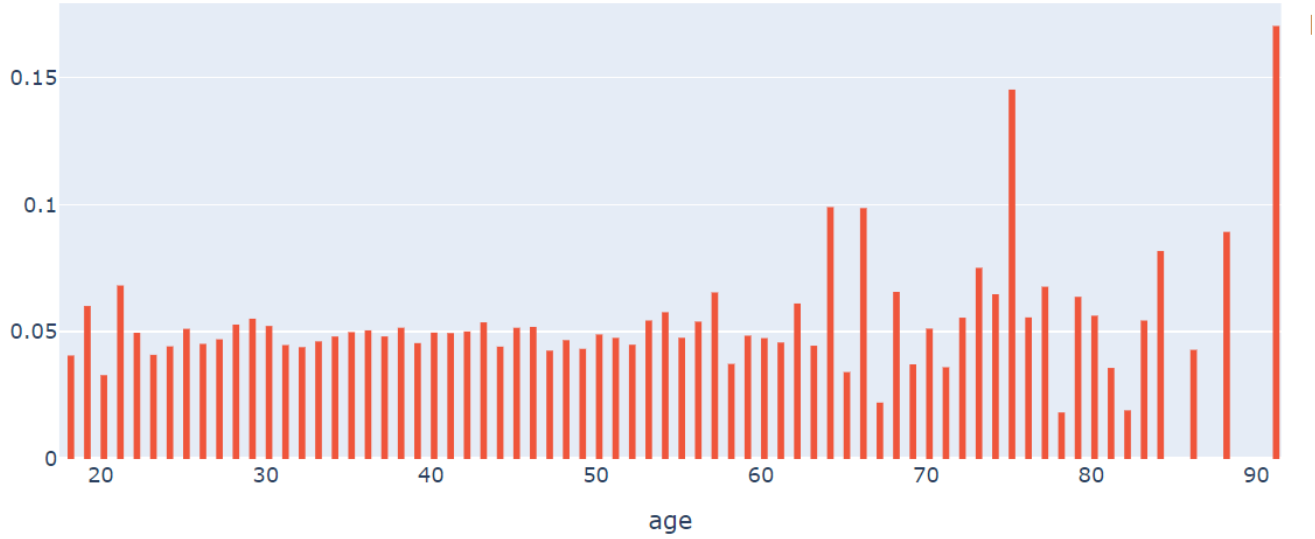

Ilustración 2 Duración vs edad con crédito

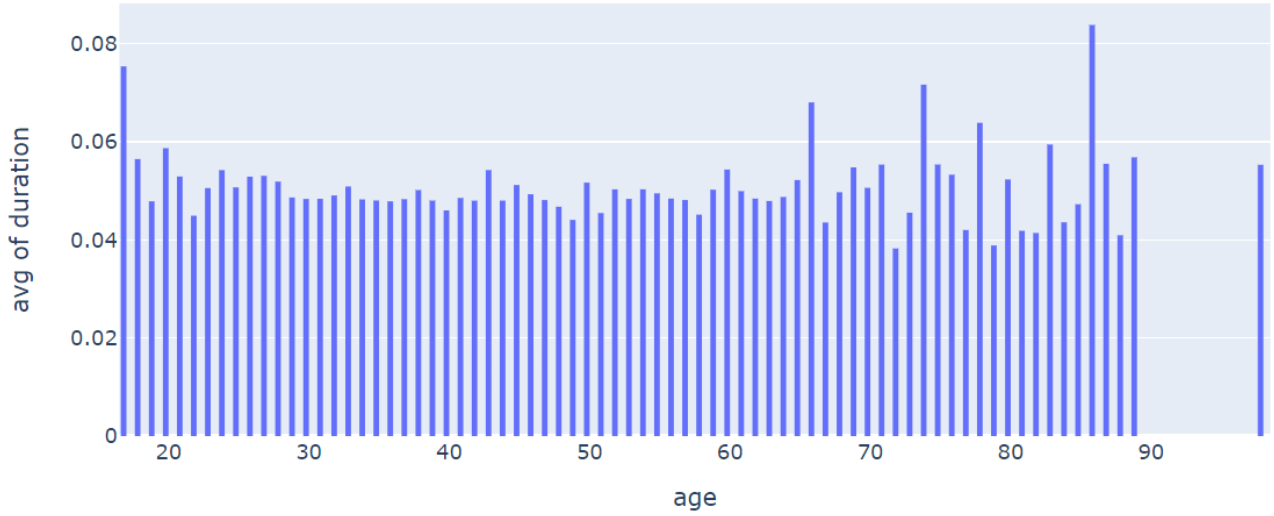

Ilustración 3 Duración vs edad sin crédito

En cuanto a la duración de la llamada dependiendo del estado civil, podemos evidenciar el hallazgo de que los estudiantes divorciados suelen demorar más en las llamadas, seguido por las personas retiradas y solteras. Esto se puede ver en la 'Ilustración 4'

Ilustración 4, Duración vs trabajo

En cuanto a la concentración de los datos de las edades, podemos ver en la 'Ilustración 5' una cantidad grande de personas entre las 30 y los 50 años que tienen un promedio de duración de 0.05 las cuales tienen crédito de vivienda, al igual que las que no tienen crédito de vivienda por lo cual este atributo no influye en la duración de llamada.

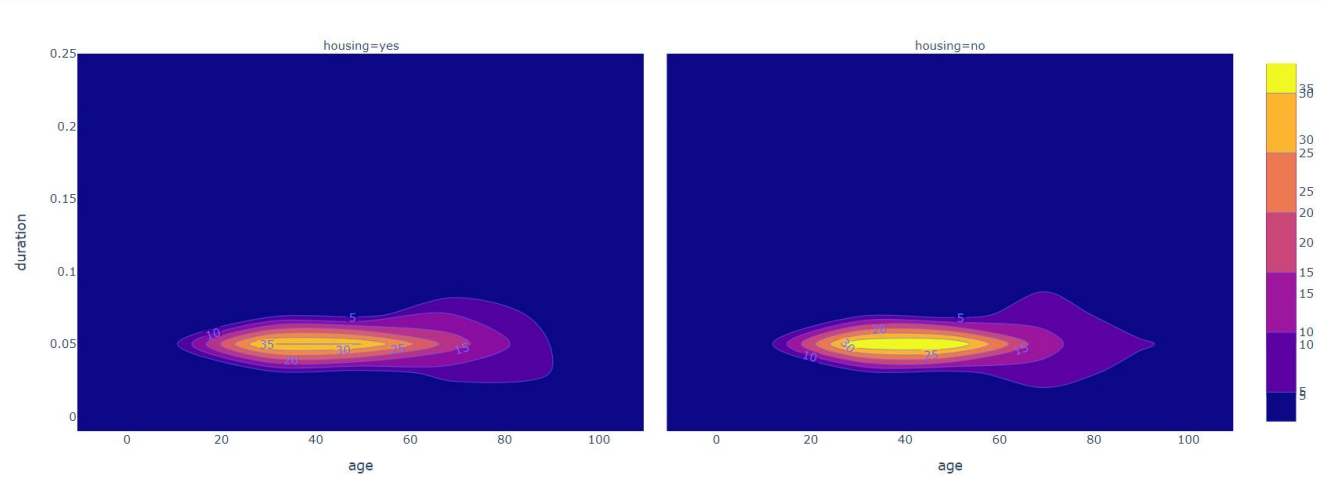

Ilustración 5 Edad vs duración

El grafico de edad vs frecuencia, nos muestra la cantidad de datos que se repiten más, o que son más frecuentes en el dataset, en este grafico podemos observar que la mayoría de los clientes se encuentran en un rango de edad entre los 25 a los 36 años, esto nos muestra claramente que los principales candidatos para adquirir una suscripción son adultos.

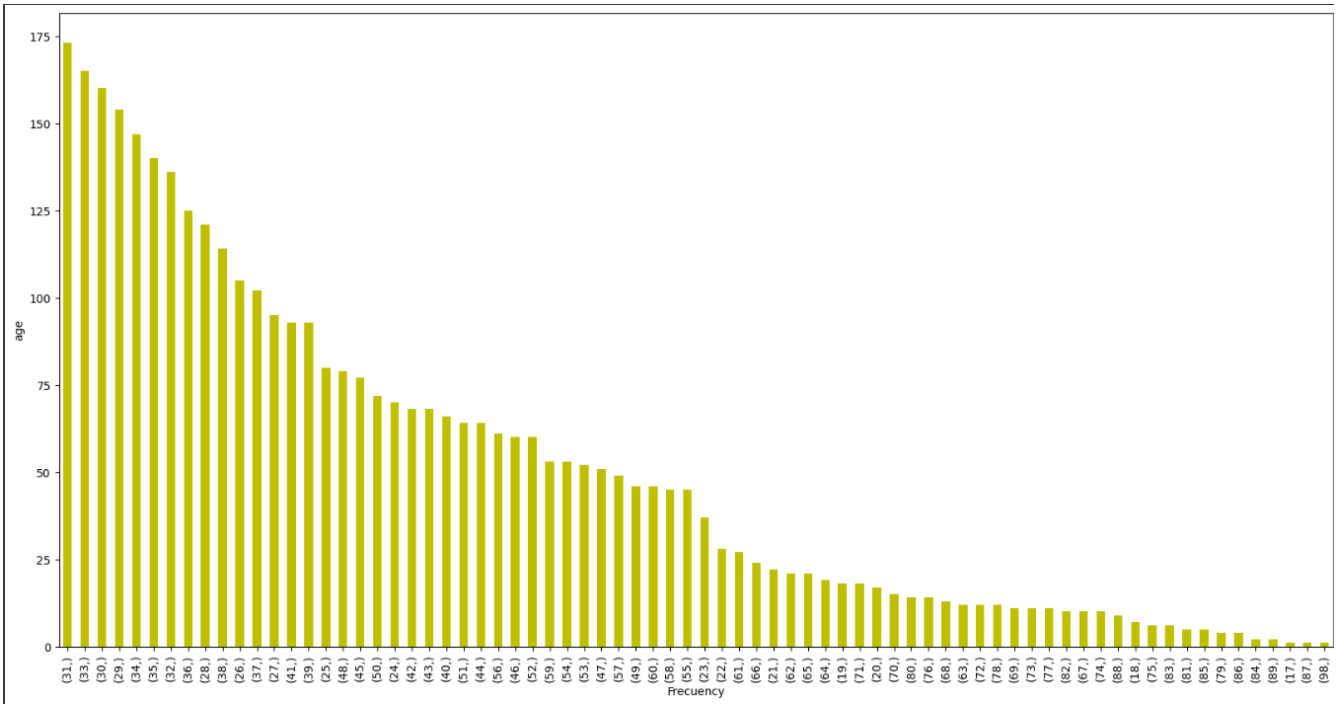

Ilustración 6, Edad vs frecuencia

En la ilustración 7 podemos ver que por medio del contacto celular duración de llamada es menor a la de contacto por teléfono, además se puede ver que la duración de llamada tiende a bajar a medida que ha tenido más contactos previos.

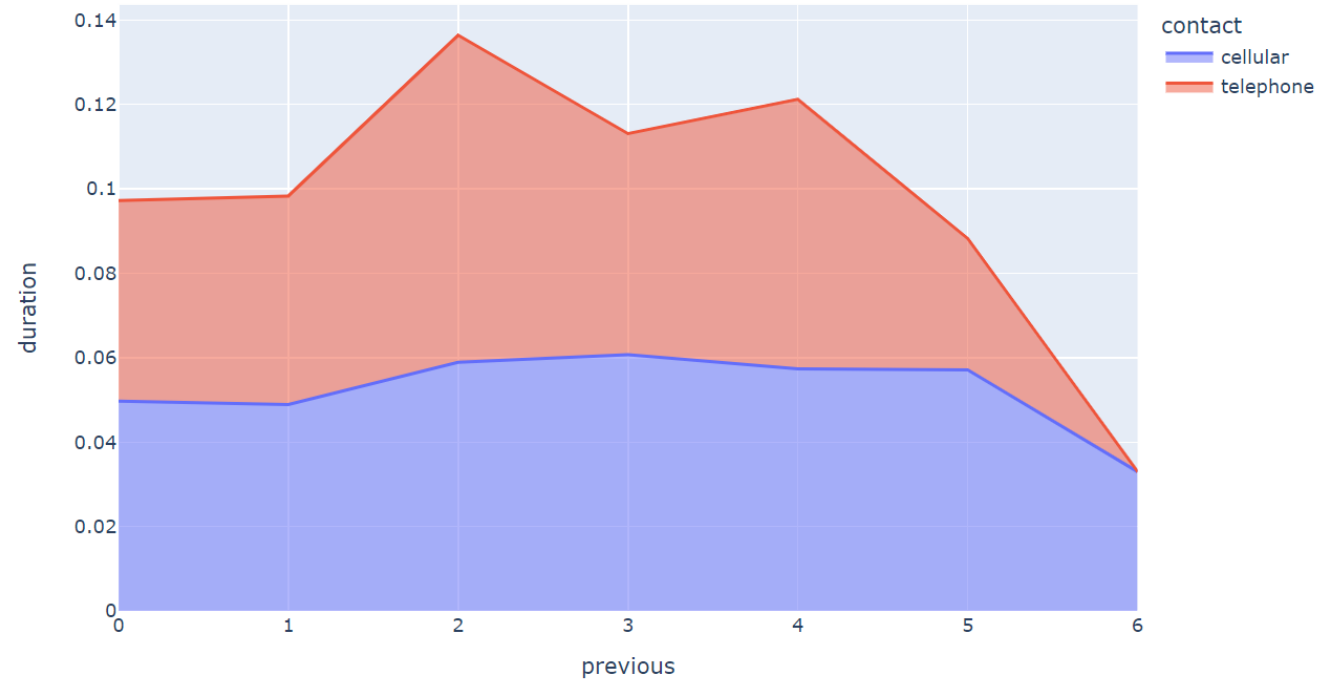

Ilustración 7, Duración vs Numeró de contactos previos

Modelos de clasificación

Al correr cada uno de los modelos para predecir la suscripción de los clientes, se evaluaron las métricas de precisión, recall, f1 score y support de cada clase, además el accuracy total y el total de aciertos y errores.

1. Método Gaussiano

Accuracy: 0.871

	precision	recall	f1-score	support
no	0.93	0.92	0.93	8781
yes	0.43	0.47	0.45	1104
accuracy			0.87	9885
macro avg	0.68	0.70	0.69	9885
weighted avg	0.88	0.87	0.87	9885

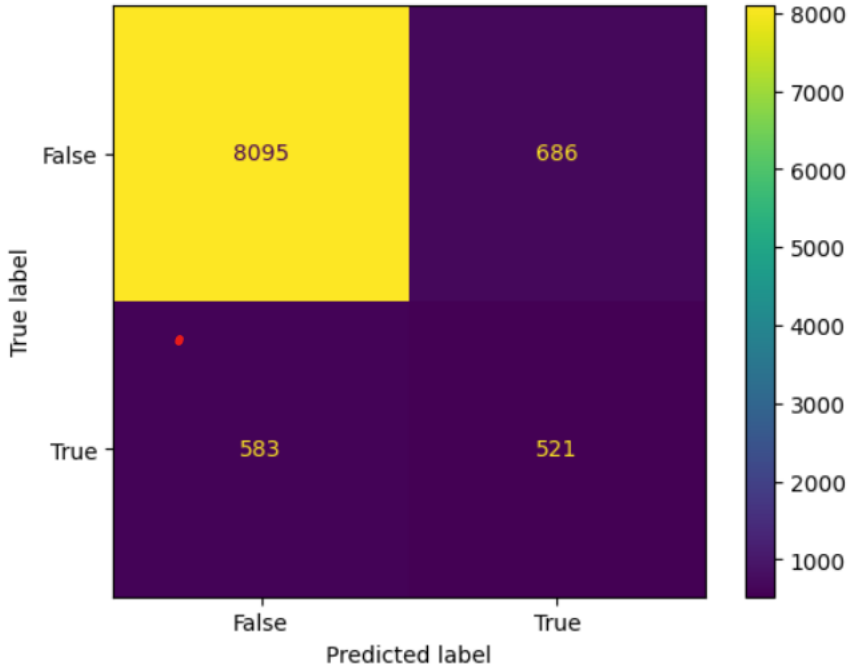

Ilustración 8 GaussNB

Como se evidencia en la 'Ilustración 8' el accuracy del modelo gaussiano de bayes fue del 87% con 8095 aciertos en que no se suscribirá y 521 aciertos en no.

2. Método Multinomial

Accuracy: 0.874

	precision	recall	f1-score	support
no	0.92	0.94	0.93	8781
yes	0.42	0.32	0.36	1104
accuracy			0.87	9885
macro avg	0.67	0.63	0.65	9885
weighted avg	0.86	0.87	0.87	9885

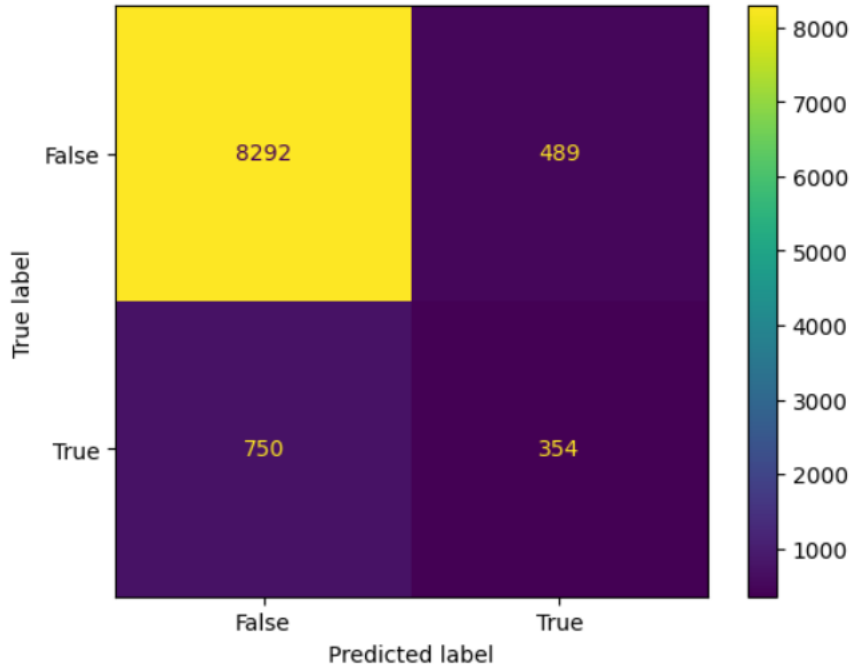

Ilustración 9 MultinomialNB

En el método multinomial como se ve en la 'Ilustración 9' se obtuvo un accuracy del 87% con 8292 aciertos en no y 354 en sí.

3. Método de Bernoulli:

Accuracy: 0.875

	precision	recall	f1-score	support
no	0.92	0.95	0.93	8781
yes	0.42	0.32	0.37	1104
accuracy			0.88	9885
macro avg	0.67	0.63	0.65	9885
weighted avg	0.86	0.88	0.87	9885

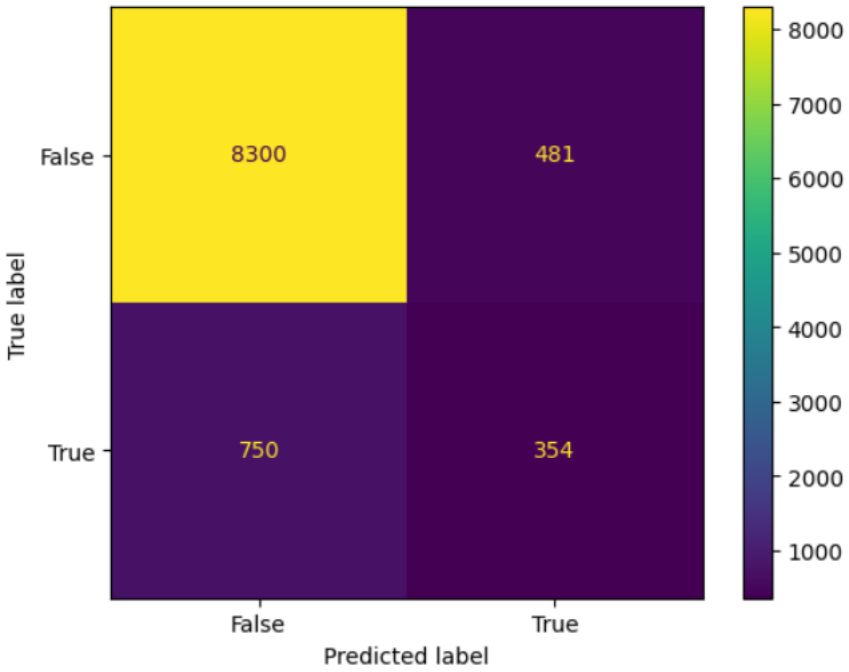

Ilustración 10 Bernoulli NB

Al ver la 'Ilustración 10'con los resultados de Bernoulli, podemos ver un accuracy de 88% con 8300 aciertos al no y 354 aciertos al sí.

-KNN Classifier

Accuracy=0.897

	precision	recall	f1-score	support
no	0.91	0.98	0.94	8781
yes	0.62	0.21	0.32	1104
accuracy			0.90	9885
macro avg	0.77	0.60	0.63	9885
weighted avg	0.88	0.90	0.87	9885

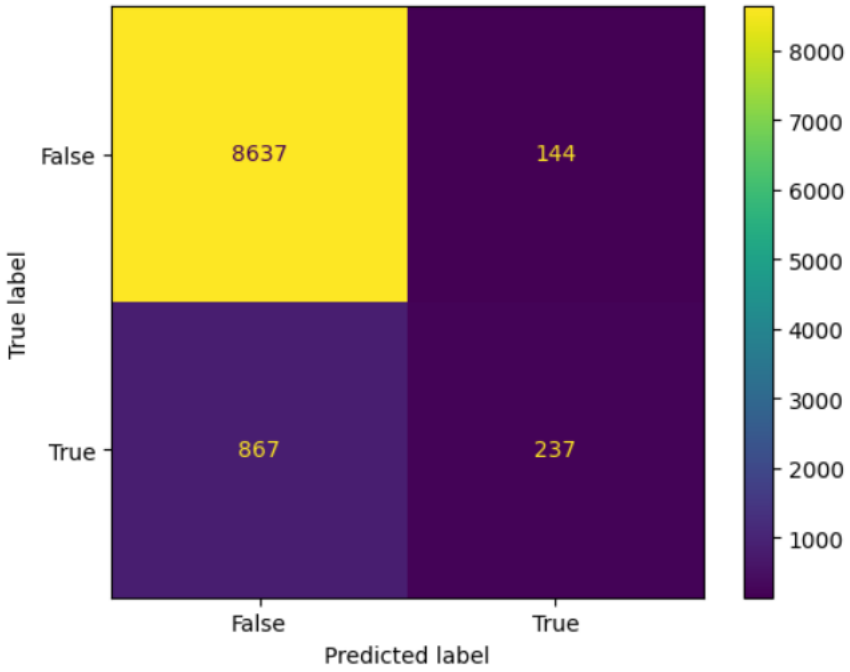

Ilustración 11 KNN

El método de KNN classifier está relacionado con el método imputación KNN en el cual se analizan los vecinos más cercanos a un valor, para poder predecir otros a partir de estos, sin embargo, en este caso funciona como un método de clasificación el cual nos permite predecir el valor de la variable dependiente. En este caso al aplicar el método, como lo vemos en la 'Ilustración 11', y hacer uso de grid cross-validation se obtuvieron los mejores parámetros para el KNN classifier y se obtuvo un accuracy de 89.76% con 8637 aciertos al no y 237 aciertos al sí.

La mejor combinación de parámetros fue el algoritmo 'ball tree' y 15 vecinos.

-Support vector machine

Accuracy= 0.90

	precision	recall	f1-score	support
no	0.91	0.99	0.95	8781
yes	0.67	0.21	0.32	1104
accuracy			0.90	9885
macro avg	0.79	0.60	0.64	9885
weighted avg	0.88	0.90	0.88	9885

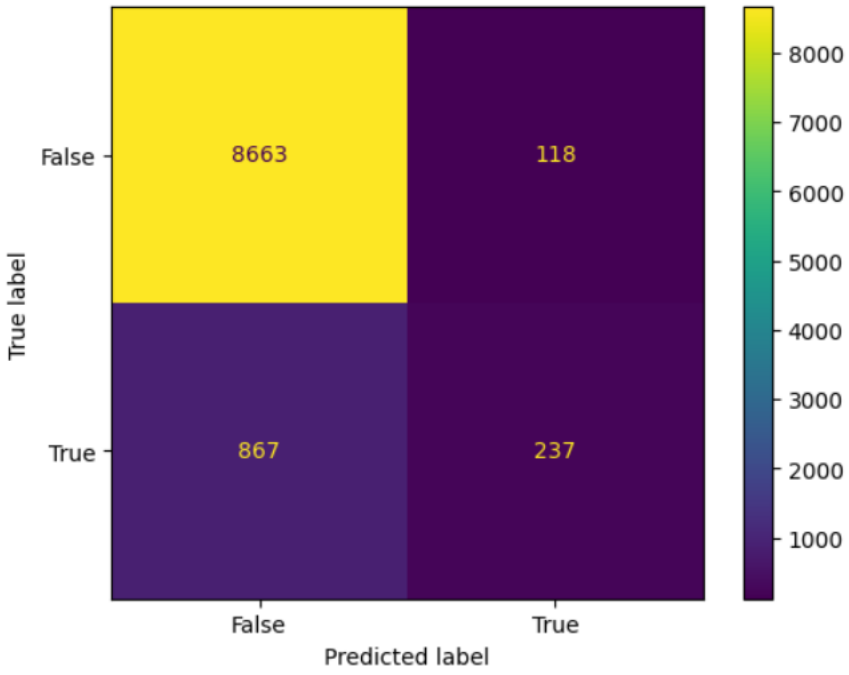

Ilustración 12 SVM

Support vector machine es un método de clasificación en el cual busca un vector con un margen más aproximado a la clasificación, a partir de unos hiperparametros los cuales ajustan de mejor manera la posición de este vector que realmente es un hiperplano, el cual separa los datos y ayuda a clasificarlos. Al hacer uso de este método, como lo vemos en la 'Ilustración 12', se aplicó un proceso de optimización para encontrar los mejores parámetros del hiperplano, este proceso de optimización es el grid cross validation; al haber encontrado los mejores parámetros se obtuvo un accuracy de 90.04% con 8663 aciertos al no y 237 aciertos al sí.

La mejor combinación de parámetros fue C=100 y Gamma=0.001.

Clustering

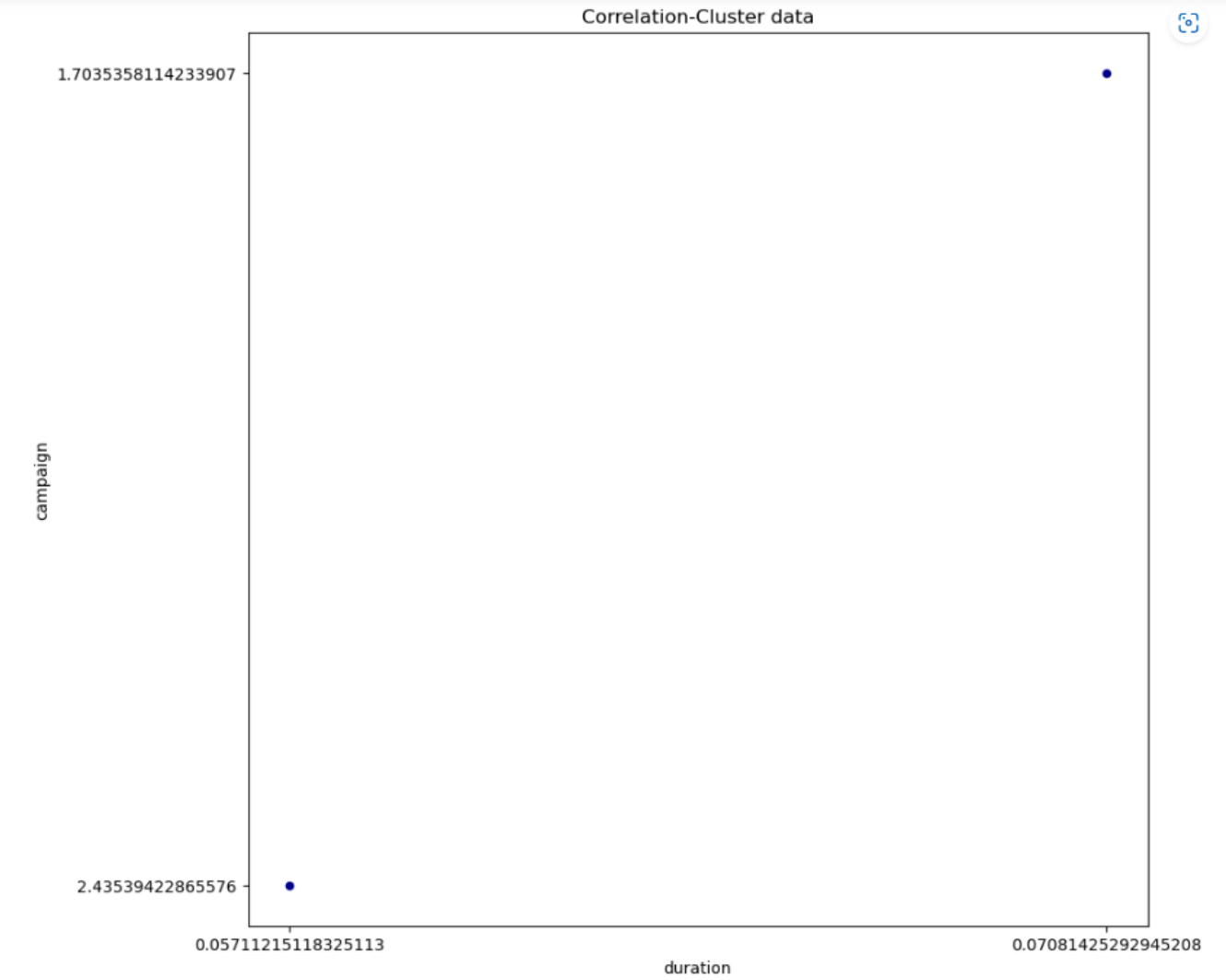

Ilustración 13, Centroides

En este caso al realizar el método de clustering, encontramos los centroides a partir de los cuales se forman los dos clusters, debido a que la información que queremos de los clientes es saber que tan posible es que estos se suscriban a un depósito a largo plazo. En este caso al aplicar el modelo de clustering encontramos que las características que más tienen relevancia para cada uno de los clusters tienen que ver con la duración del tiempo en la que se interactuó con el cliente y el número de veces que se contactó con estos.

4.4 Discusión

Haciendo un análisis de todos los métodos, utilizados, podemos evidenciar que el modelo que tuvo más exactitud fue el Support vector machine con 90% de accuracy, seguido de K-NN, y naive bayes, como lo vemos en la 'Tabla 2', sin embargo, los modelos tuvieron buenas métricas, teniendo en cuenta que los valores se les aplicó la técnica de cross-validation, sus resultados fueron mejores aplicando los mejores parámetros para SVM y K-NN, sin embargo, la exactitud es fiable cuando se cuenta con un dataframe balanceado, lo que quiere decir que tiene igual cantidad de cada clase en el dataframe, lo cual en nuestro caso no se cumple, debido a que se tiene más cantidad de la clase 'No' que de la clase 'Si', esto podemos verlo simplemente viendo la matriz de confusión, si vemos la de Support vector machine en la 'Ilustración 12, vemos 8663 aciertos al 'No', 118 errores contra 237 aciertos al 'Si' y 867 errores, lo que nos hace ver que la clase predominante es el 'No' con más del 88% de los datos, a causa de esto debemos ver otra clase de métricas [14].

Modelo	Parametros	Accuracy
SVM	C=100;Gamma=0.001	90%
K-NN	Algorith='ball_tree; Neigh=15	89,7%
Naive bayes Bernoulli	Defl	87,5%
Naive bayes Gaussian	Defl	87,1%
Naive bayes Multinomial	Defl	87,4%

Tabla 2 Accuracy

Haciendo un análisis del F1score podemos ver que los resultados cambian teniendo el modelo de naive bayes con el mejor F1 score del 45% para la clase 'Si', y 93% para la clase 'No', esto sumado a la exactitud encontrada del 87,1% sería el algoritmo más apto para la predicción de si las personas se suscribirían al producto del deposito o no, debido a que de nada sirve un algoritmo que tenga una excelente exactitud que acierte a todas las clases que 'No' se suscriban al producto y erre a todas las clases donde 'Si' se suscribirían, ya que el objetivo de este algoritmo es encontrar personas que realmente si se suscribirían al producto para incrementar los ingresos.

Teniendo en cuenta además el análisis exploratorio de datos, podemos evidenciar resultados bastante válidos para encontrar las personas que más se suscribirían al producto. Al tener en cuenta que la duración de las llamadas es muy importante, apreciando que se trata de una campaña de llamadas, se puede llegar a conclusiones como que las personas mayores tienden a durar más tiempo en llamada, que los datos se concentran entre personas entre los 39 a 59 años, que ocupación y estado civil tienen las personas que se suscriben, que tipo de productos tienen abiertos los clientes y que tipo de comunicación se utilizó, lo cual nos ayuda a encontrar un segmento y un foco de personas que se suscribirían al programa de manera que se puede filtrar y el número de suscripciones aumente.

En cuanto a las nuevas tecnologías para el mercado bancario, este ejercicio y los modelos implementados para predecir las personas que se suscribirían al producto, ejemplifica la forma en que las bases de datos, sus análisis y modelos cambian, mejoran e impactan la experiencia de los consumidores, las estrategias y ganancias, debido a que con análisis de datos no solo podemos predecir quien se suscribiría a un producto, sino que también se puede calcular el riesgo de que una persona pueda pagar un crédito o no, se calcula que personas adquirirían una tarjeta de crédito con distintas características de viajes o compras dependiendo de su estilo de vida, el cual se comparte día a día por el uso de redes sociales y como lo podemos ver en la literatura del nuevo Digital banking y experiencia del consumidor de UK bank managers [15].

Al analizar los resultados del modelo podremos observar que entre mayor es el tiempo y la cantidad de veces que se interactúa con cliente, es más probable que este se suscriba y se obtenga una salida exitosa en el proceso de marketing (véase el dataframe de los centroides en el archivo ipynb).

5 Conclusión

Dentro del articulo se desarrolló el concepto de nuevas tecnologías dentro del mercado bancario, enfocado en la minería de datos, aplicando análisis exploratorio de datos donde encontramos varias claves de segmentación del mercado como edad, tipo de llamada, ocupación, estado civil y demás, filtrando el tipo de persona analizada.

Posteriormente vimos 3 métodos de clasificación no supervisada de los cuales primó el Support vector machine con mejor exactitud, sin embargo, pudimos evidenciar que el dataframe estaba desbalanceado en sus clases, evaluando el F1 score, el modelo más exacto, balanceado y preciso fue el naive bayes gaussiano concluyendo que para modelos balanceados es mejor Support Vector machine con técnica de mejoramiento cross-validation y para modelos desbalanceados es mejor utilizar naive bayes en el entrenamiento.

Además, se realizó un clustering para analizar que este proceso nos ayuda a determinar que entre más tiempo se interactúa con el cliente, mayor probabilidad se tiene de que se suscriba al programa.

Por último, se puede decir que la tecnología de machine learning y análisis exploratorio de datos, ramas de la minería de datos, están cambiando, e impactando el mercado bancario y no solo este, sino todas las áreas de estudio como lo vimos con la literatura, enfocando modelos a la meteorología, la medicina y la construcción.

6 Bilbiography

[1] Florin Gorunescu, *Data Mining Concepts, Models and Techniques*. Springer, 2009.

[2] V. A. Brei, "Machine learning in marketing," *Foundations and Trends in Marketing*, vol. 14, no. 3, pp. 173–236, Aug. 2020, doi: 10.1561/1700000065.

[3] C. I. Mbama, P. Ezepue, L. Alboul, and M. Beer, "Digital banking, customer experience and financial performance: UK bank managers' perceptions," *Journal of Research in Interactive Marketing*, vol. 12, no. 4, pp. 432–451, 2018, doi: 10.1108/JRIM-01-2018-0026.

[4] S. Radovanović, A. Petrović, B. Delibašić, and M. Suknović, "A fair classifier chain for multi-label bank marketing strategy classification," *International Transactions in Operational Research*, vol. 30, no. 3, pp. 1320–1339, 2023, doi: 10.1111/itor.13059.

[5] Di. Cook, E. K. Lee, and M. Majumder, "Data Visualization and Statistical Graphics in Big Data Analysis," *Annual Review of Statistics and Its Application*, vol. 3. Annual Reviews Inc., pp. 133–159, Jun. 01, 2016. doi: 10.1146/annurev-statistics-041715-033420.

[6] J.-H. Chen, "KNN based knowledge-sharing model for severe change order disputes in construction," *Autom Constr*, vol. 17, no. 6, pp. 773–779, 2008, doi: 10.1016/j.autcon.2008.02.005.

[7] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using Naïve Bayes Networks," in *2017 6th International Conference on Systems and Control, ICSC 2017*, 2017, pp. 426–432. doi: 10.1109/ICoSC.2017.7958686.

[8] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," *Soft Computing Letters*, vol. 3, p. 100014, Dec. 2021, doi: 10.1016/J.SOCL.2021.100014.

[9] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, doi: 10.1016/j.neucom.2013.04.011.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[11] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," *Soft Computing Letters*, vol. 3, p. 100014, Dec. 2021, doi: 10.1016/J.SOCL.2021.100014.

[12] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using Naïve Bayes Networks," in *2017 6th International Conference on Systems and Control, ICSC 2017*, 2017, pp. 426–432. doi: 10.1109/ICoSC.2017.7958686.

[13] J.-H. Chen, "KNN based knowledge-sharing model for severe change order disputes in construction," *Autom Constr*, vol. 17, no. 6, pp. 773–779, 2008, doi: 10.1016/j.autcon.2008.02.005.

[14] Florin Gorunescu, *Data Mining Concepts, Models and Techniques*. Springer, 2009.

[15] C. I. Mbama, P. Ezepue, L. Alboul, and M. Beer, "Digital banking, customer experience and financial performance: UK bank managers' perceptions," *Journal of Research in Interactive Marketing*, vol. 12, no. 4, pp. 432–451, 2018, doi: 10.1108/JRIM-01-2018-0026.