Tecnológico de Monterrey

Análisis y Reporte sobre el desempeño del modelo.

Luis Gerardo Lagunes Najera A01272515

Septiembre 2023

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Prof. Jorge Adolfo Ramírez Uresti

Análisis y Reporte sobre el desempeño del modelo.

Para la selección de los datos de las correspondientes practicas se uso la pagina web Kaggle para poder encontrar un data set idóneo para la actividad por lo cuál tenía que contener alguna variable a predecir y diversas variables para entrenamiento que podrían ser categóricas o numéricas, en este caso eran de ambos tipos de datos, tanto numéricas como categóricas.

Este data set contiene 21 columnas con diferentes tipos de datos, en el caso encontrado en este data set se pretende predecir si una persona de acuerdo con diversos estatus puede o no estar endeudada con el banco, esto podría servir para poder evaluar a una persona cuando va a adquirir algún préstamo o crédito y analizar si implica una pérdida o ganancia en el banco a analizar.

Antes de todo esto se procedió a el preprocesamiento de los datos ya que el data set contenía algunos valores innecesarios para el análisis como por ejemplo el número de teléfono, el día o mes que iban al banco, al igual que se procedió a la eliminación de data vacía, ya que esta podría alterar los resultados generados por nuestros modelos debido a valores vacíos. Al igual que se procedió a categorizar las edades para poder obtener una mejor predicción de acuerdo con grupo de edades.

	age	job	marital	education
0	41-60	housemaid	married	basic.4y
2	19-40	services	married	high.school
3	41-60	admin.	married	basic.6y
4	41-60	services	married	high.school
6	41-60	admin.	married	professional.course
...	...	...	...	...
41183	61-80	retired	married	professional.course
41184	41-60	blue-collar	married	professional.course
41185	41-60	retired	married	university.degree
41186	41-60	technician	married	professional.course
41187	61-80	retired	married	professional.course

30488 rows × 21 columns

1.- Preprocesamiento de datos con Python.

Una vez teniendo los datos limpios, se hizo una elección de datos de predicción, es decir se tomaron 4,000 muestras de el total de datos, de este total de datos 2,000 tienen valor de “yes” en la variable a predecir y las otras 2,000 el valor de “no”, esto con la finalidad de tener los datos a una buena proporción y no ocasionar un modelo sesgado en donde no se tomen en cuenta en los datos de prueba o entrenamiento alguno de los valores a predecir.

```
si_samples = data_limpia[data_limpia['housing'] == 'yes'].sample(n=2000)
no_samples = data_limpia[data_limpia['housing'] == 'no'].sample(n=2000)
```

2.- Elección de los datos

Una vez teniendo los 4,000 datos solo seleccionamos las variables más significativas y que contengan información importante de nuestro data set, en este caso las variables que usaremos para predicción serán: 'age', 'job', 'marital', 'housing', 'contact', 'month', 'day_of_week', 'previous'. Y nuestra variable a predecir será: ‘loan’.

Por consiguiente, definimos así nuestros valores de “y” y de “X” para después hacer la separación y evaluación del modelo.

```
Y = data_limpia['loan'].values
X = data_limpia[['age', 'job', 'marital', 'housing', 'contact',
                 'month', 'day_of_week',
                 'previous']].values
```

3.- Selección de X y Y

Ahora bien, se separaron los datos en datos de entrenamiento y de test con la librería de sklearn usando la función *train_test_split* en donde solo el 5% de los datos iban a ser seleccionados para hacer las pruebas y el resto de entrenamiento, estos datos iban a cambiar en cada iteración realizada en nuestro ciclo para así obtener diversos resultados de acuerdo a diversos datos tanto de entrenamiento como de prueba.

```
========================================
------ X_train --------
[[ 1  9  2 ...  3  4  0]
 [ 1  0  1 ...  3  3  0]
 [ 1 10  1 ...  7  2  0]
 ...
 [ 1  0  1 ...  1  3  0]
 [ 0  9  1 ...  3  4  0]
 [ 0  6  2 ...  6  3  0]]
------------------------------------------------------------
```

4.- X_train en la iteración 5

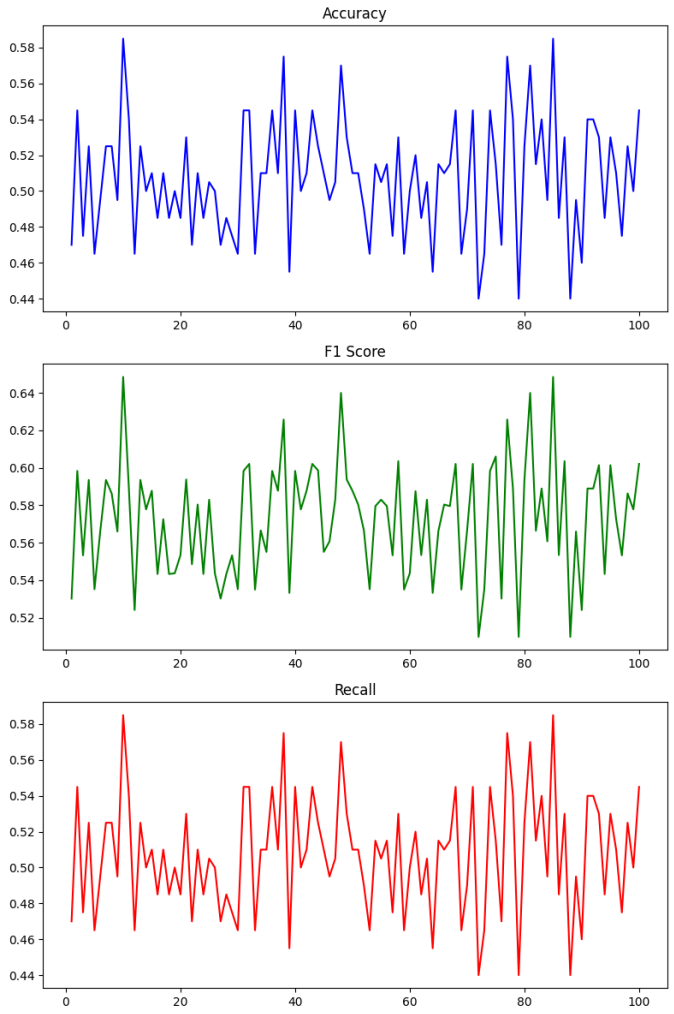
```
------ y_train --------
['no' 'yes' 'no' ... 'no' 'yes' 'yes']
-----------------------------------------------
```

5.- y_train en la iteración 5

```
------ X_test --------
[[0 7 1 ... 6 3 0]
 [0 4 2 ... 6 3 0]
 [0 6 2 ... 0 2 0]
 ...
 [0 9 2 ... 4 2 0]
 [1 9 1 ... 4 2 0]
 [0 9 1 ... 1 0 0]]
----------------------------------------------------------------
------ y_test --------
['no' 'no' 'no' 'no' 'yes' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'yes' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'no'
 'no' 'yes' 'no' 'no' 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes'
 'no' 'no' 'no' 'no' 'yes' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'yes' 'yes' 'yes' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'no'
 'no' 'no' 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'yes' 'yes' 'yes' 'no'
 'no' 'no' 'no' 'yes' 'no' 'no' 'no' 'no' 'no' 'no' 'no' 'yes' 'no' 'yes'
 'no' 'no' 'no' 'yes']
----------------------------------------------------------------
```

Una vez se tiene todo esto se procese a aplicar el algoritmo Naive Bayes Bernulli debido a que este modelo se ajusta mejor con los datos debido a las características que estos poseen, se evalúa con un modelo simple aplicando la librería de BernulliNB(), después de esto se hacen 100 iteraciones con este modelo cambiando de datos y probando el modelo sin aplicar ajustes en los parámetros para ajustar el modelo, a lo cuál se llega a un modelo con unas métricas bastante bajas, analizando las varianzas entre las métricas generadas se obtienen varianzas en la Accuracy de 0.0011, la varianza en el F1 Score de 0.0009 y una varianza en el Recall de 0.0011, aunque son números bastante pequeños podemos ver que el modelo sigue siendo deficiente a lo esperado pero muestra que no hay sesgo a los datos debido a que las métricas de Accuracy, F1 Score y Recall en las 100 iteraciones son bastante consistentes y tienen una baja variabilidad. Lo que nos quiere decir que el modelo tiende a producir resultados muy similares en diferentes divisiones de datos aleatorias o configuraciones de entrenamiento.

Al igual que esto mismo puede ser indicativo de que el modelo generaliza bien a datos nuevos y que no está sobre ajustando (overfitting) o siendo demasiado sensible a las pequeñas variaciones en los datos de entrenamiento.



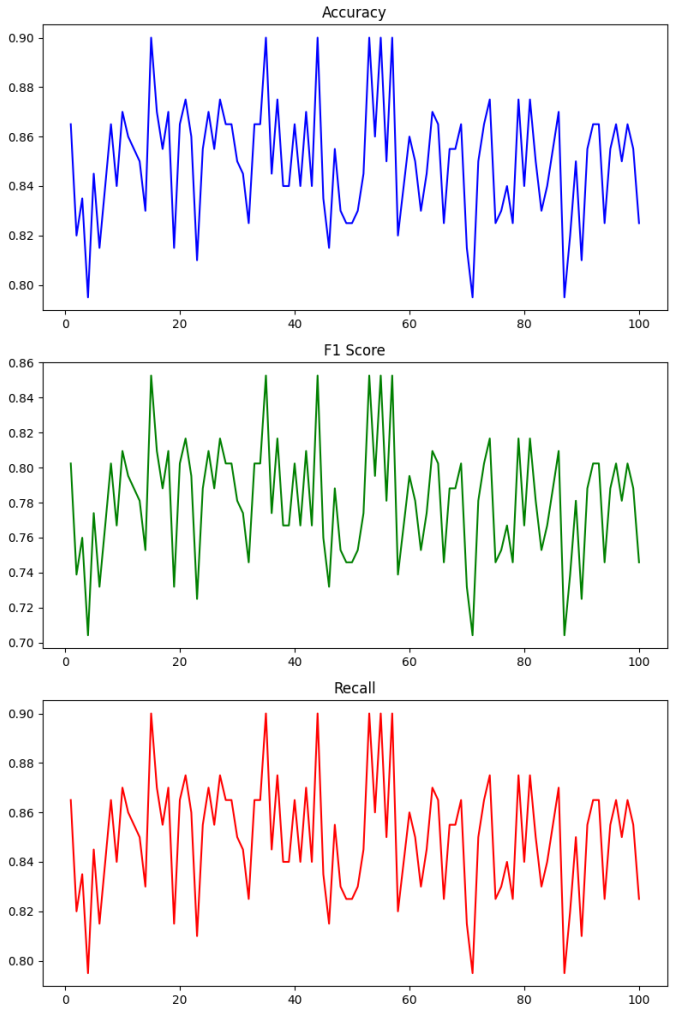
7.- Gráfico de modelo sin ajustes ni modificaciones.

```
Varianza de Accuracy: 0.0011
Varianza de F1 Score: 0.0009
Varianza de Recall: 0.0011
```

8.- Valores de varianza entre métricas

Una vez teniendo un modelo que no muestra overfitting, ni sesgo en los datos, ni un grado de varianza malo en las métricas se procedió a ajustar los parámetros para mejorar el desempeño del modelo, en este caso se utilizo un suavizado de 0.5 con el cuál se obtenían valores más contundentes en los datos, utilizamos el fit_prior como verdadero ya que esto hace que el modelo aprenda de las probabilidades previas para entregar una mejor predicción al igual que el set de test se valoro en el 5% de los datos, una vez aplicados estos nuevos parámetros en el modelo se volvió a evaluar garantizando mejores resultado en las métricas teniendo así varianzas aún más bajas que las del primer modelo en las métricas.

Lo cuál nos indica que el modelo tiene una buena parametrización alcanzando accuracys y recalls en algunos puntos arribas del 0.90 al igual que un F1 Score arriba del 0.80 lo cuál indica que los ajustes de los parametros mejoraron el modelo de manera significativa y mostrando una varianza baja lo cuál indica que tiene una baja variabilidad lo cuál indica que los resultados son coherentes y predecibles.



9.-Gráfico del modelo con ajustes

```
Varianza de Accuracy: 0.0005
Varianza de F1 Score: 0.0011
Varianza de Recall: 0.0005
```

10.- Varianzas de cada métrica

Aquí una comparación final de los dos modelos evaluados: