Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

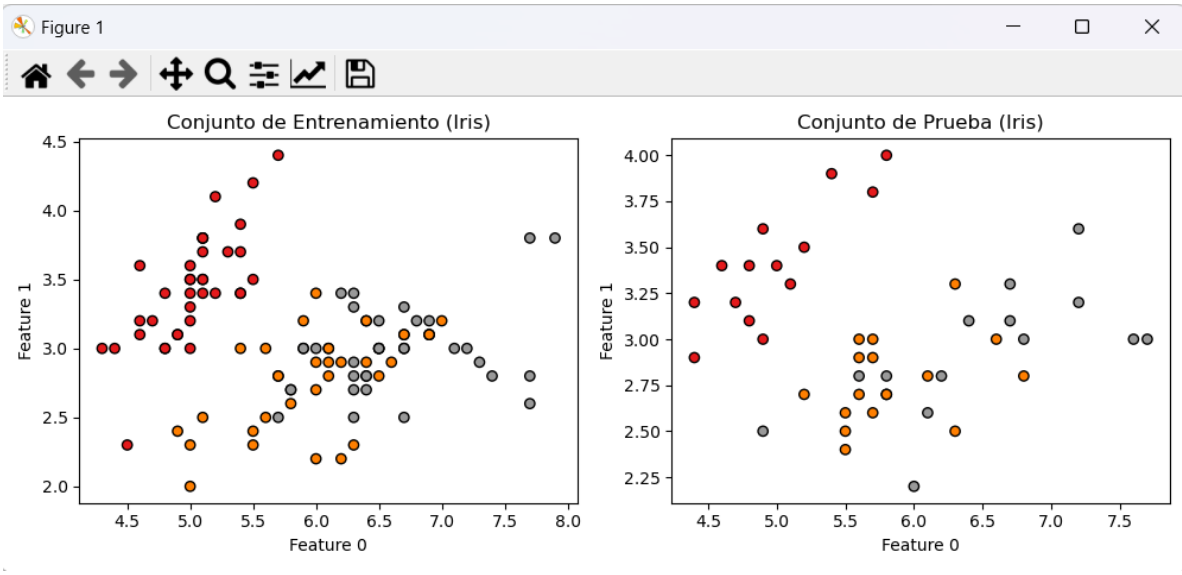
Septiembre, 2023

Análisis y Reporte sobre el desempeño del modelo

Jorge Isidro Blanco Martínez - A01745907

En el presente reporte se presentará un análisis de desempeño acerca de un modelo knn implementado en Python con el uso de la librería sklearn mediante el uso de gráficos generados en Python con matpoltlib y seaborn

Separación y evaluación del modelo con un conjunto de prueba y un conjunto de entrenamiento



Para la prueba de este modelo knn se utilizó el dataset iris de la librería sklearn ya que es un dataset de clasificación lo cual es preferible para utilizar en un modelo como knn, el dataset fue dividido 70% para el conjunto de entrenamiento y 30% para validación y pruebas, en el gráfico superior podemos apreciar los datos de ambos conjuntos, cada clase está representada por un color diferente por lo que podemos comprobar que son diferentes así como la distribución de las diferentes clases además, podemos observar que ambos contienen los diferentes tipos de clases que hay en el dataset.

```
Features: [5.4 3.9 1.7 0.4], Etiqueta: 0
Features: [5.8 4.  1.2 0.2], Etiqueta: 0
Features: [5.1 3.5 1.4 0.3], Etiqueta: 0
Features: [5.4 3.4 1.5 0.4], Etiqueta: 0
Features: [5.2 3.5 1.5 0.2], Etiqueta: 0
Features: [6.6 2.9 4.6 1.3], Etiqueta: 1
Features: [4.6 3.6 1.  0.2], Etiqueta: 0
Features: [5.8 2.6 4.  1.2], Etiqueta: 1
Features: [5.  2.  3.5 1. ], Etiqueta: 1
Features: [6.  2.9 4.5 1.5], Etiqueta: 1
```

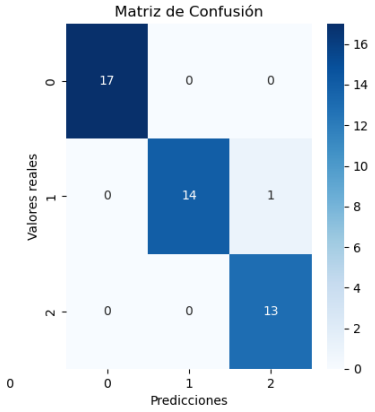
En la imagen anterior se muestran 10 datos aleatorios del dataset iris donde Features son los valores de entrada y la etiqueta es el conjunto al que pertenece

Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

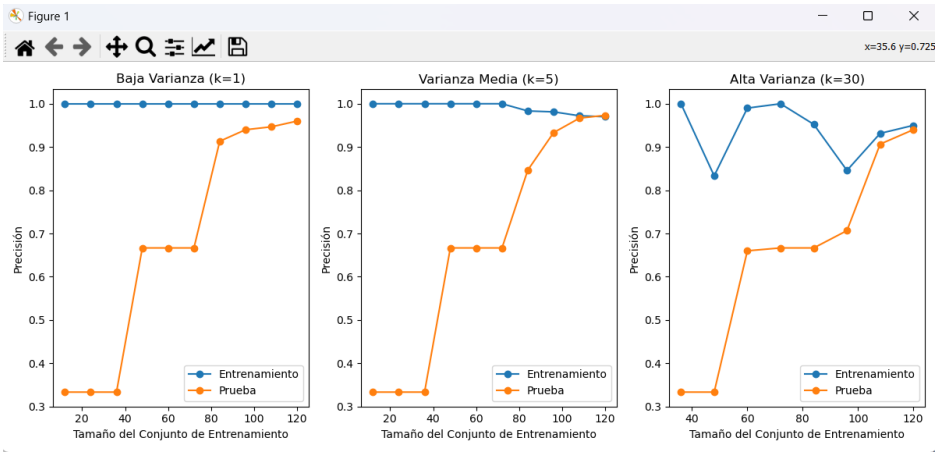
Para medir la precisión del modelo utilicé la función accuracy score de sklearn la cual permite conocer el porcentaje de acierto del modelo, esto es posible ya que estamos hablando de un modelo de aprendizaje supervisado por lo que tenemos las “Y” objetivos reales y las podemos comparar con las arrojadas por el modelo

```
k utilizada:  3
Precisión del modelo con Framework: 0.9777777777777777
```

A continuación obtenemos la matriz de confusión para conocer los resultados predichos y los reales:



Con la matriz de confusión de las pruebas podemos ver que el modelo tiene un bajo nivel de sesgo ya que tuvo 97% de precisión y solo erró en un valor de las predicciones por lo que podemos comprobar que el modelo está ajustado de buena manera



Con los gráficos anteriores podemos observar el cambio en la varianza del modelo según el ajuste de la k el modelo cambia en cuanto a precisión en el modelo knn entre menor sea la k el modelo tendrá mayor varianza ya que se producirá un overfitting por lo que al hacer predicciones puede que no sea tan eficaz como aparenta ya que es probable que solo se haya aprendido los datos. En

el primer gráfico, el modelo tiene una alta varianza y muestra una brecha significativa entre el rendimiento en entrenamiento y prueba, indicando overfitting. En el segundo gráfico (k=5), el modelo muestra una varianza moderada y un equilibrio entre el rendimiento en ambos conjuntos. En el tercer gráfico, el modelo tiene una baja varianza y un rendimiento más similar en entrenamiento y prueba, indicando underfitting.

Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt

Para encontrar el mejor hiperparametro k hacemos interpretación de los gráficos de aprendizaje anteriores para darnos cuenta que lo mejor para este modelo sería una k entre 2 y 5 ya que de esta manera no estará underfit o overfit para esto hacemos pruebas con distintos valores de k

```
k utilizada:  6
Precisión del modelo con Framework: 0.9333333333333333

k utilizada:  5
Precisión del modelo con Framework: 0.9555555555555556

k utilizada:  3
Precisión del modelo con Framework: 0.9777777777777777
```

Mediante un ciclo for que pruebe los valores de k entre 2 y 10 guardando el de mejor precisión y valor numérico más bajo obtenemos que 3 es un buen valor de k

Finalmente, con lo visto anteriormente podemos comprobar la importancia de la selección cuidadosa del conjunto de datos se identificó como un paso importante en el entrenamiento de un modelo y la aplicación de este, junto con la comprensión del sesgo y la varianza en función del valor de k. Se demostró que un k pequeño tiende a aumentar el sesgo y la varianza, lo que puede resultar en un sobreajuste, mientras que un k grande reduce la varianza, pero podría introducir sesgo. Por lo tanto, se enfatizó la importancia de ajustar apropiadamente el valor de k para encontrar un equilibrio óptimo entre sesgo y varianza, destacando así la relevancia de estas consideraciones en el desempeño y la generalización efectiva de los modelos de aprendizaje máquina.