Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo (Portafolio Análisis)

Rubi Royval Larios A01754304

ITESM CEM

Grupo: 101

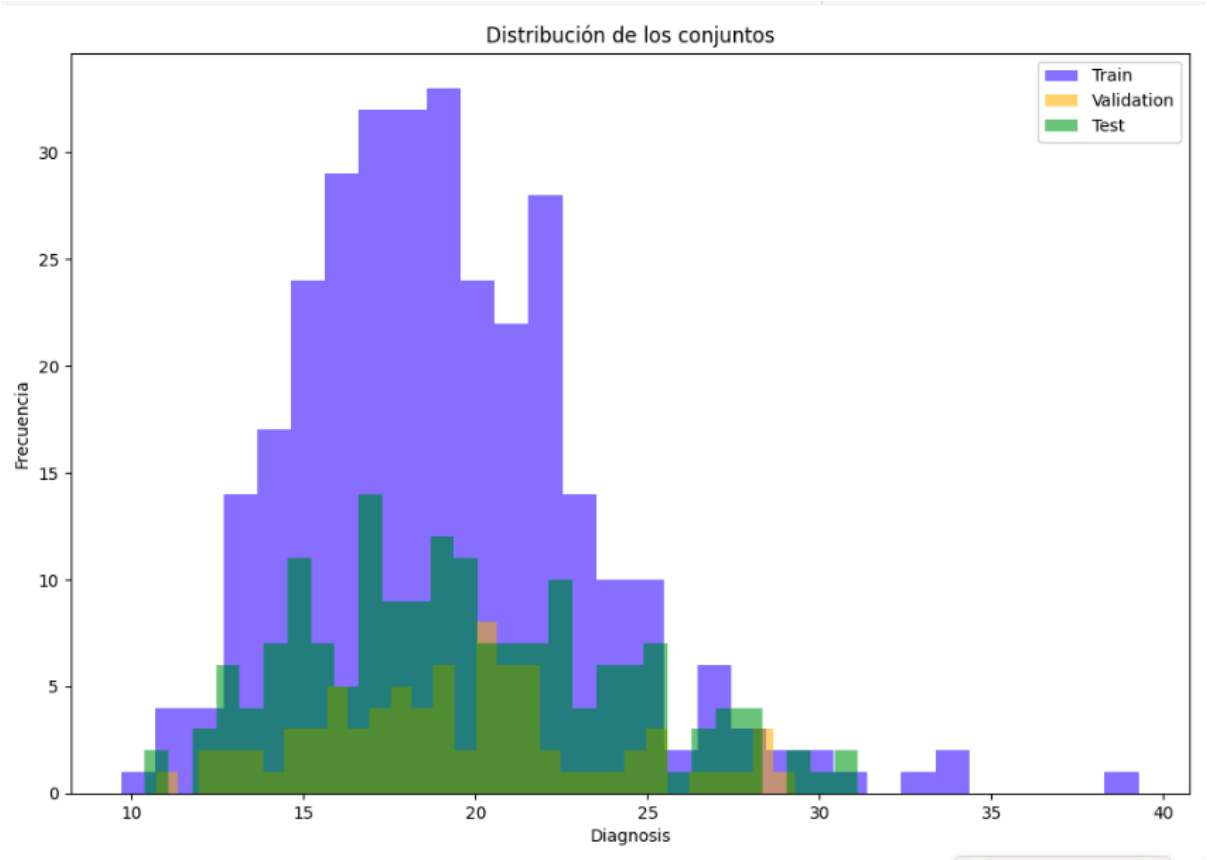
Instituto Tecnológico y de Estudios Superiores de Monterrey

Dataset

Para esta evidencia decidí utilizar el Breast Cancer Wisconsin Dataset ya que el análisis decidí hacerlo utilizando el algoritmo K-Nearest Neighbors (KNN) y esta base datos es popular para problemas de clasificación, por lo tanto es adecuada para poder realizar el análisis. Dicho dataset en general contiene diferentes variables (numéricas) las cuales llevan a decidir si un tumor es benigno o maligno.

Separación y evaluación del modelo

Para la división de los datos, del total del dataset tomamos el 70% para el conjunto de entrenamiento, el 30% del total del dataset para el conjunto de prueba, después del conjunto de entrenamiento tomamos el 20% para el conjunto de validación. Para la visualización de esta distribución decidí graficar la característica 'diagnosis' ya que es donde mejor se observaba dicha distribución, de igual forma imprimí el tamaño de los conjuntos.



```
Conjunto de Entrenamiento:
(318,)

Conjunto de Validación:
(171,)

Conjunto de Prueba:
(80,)
```

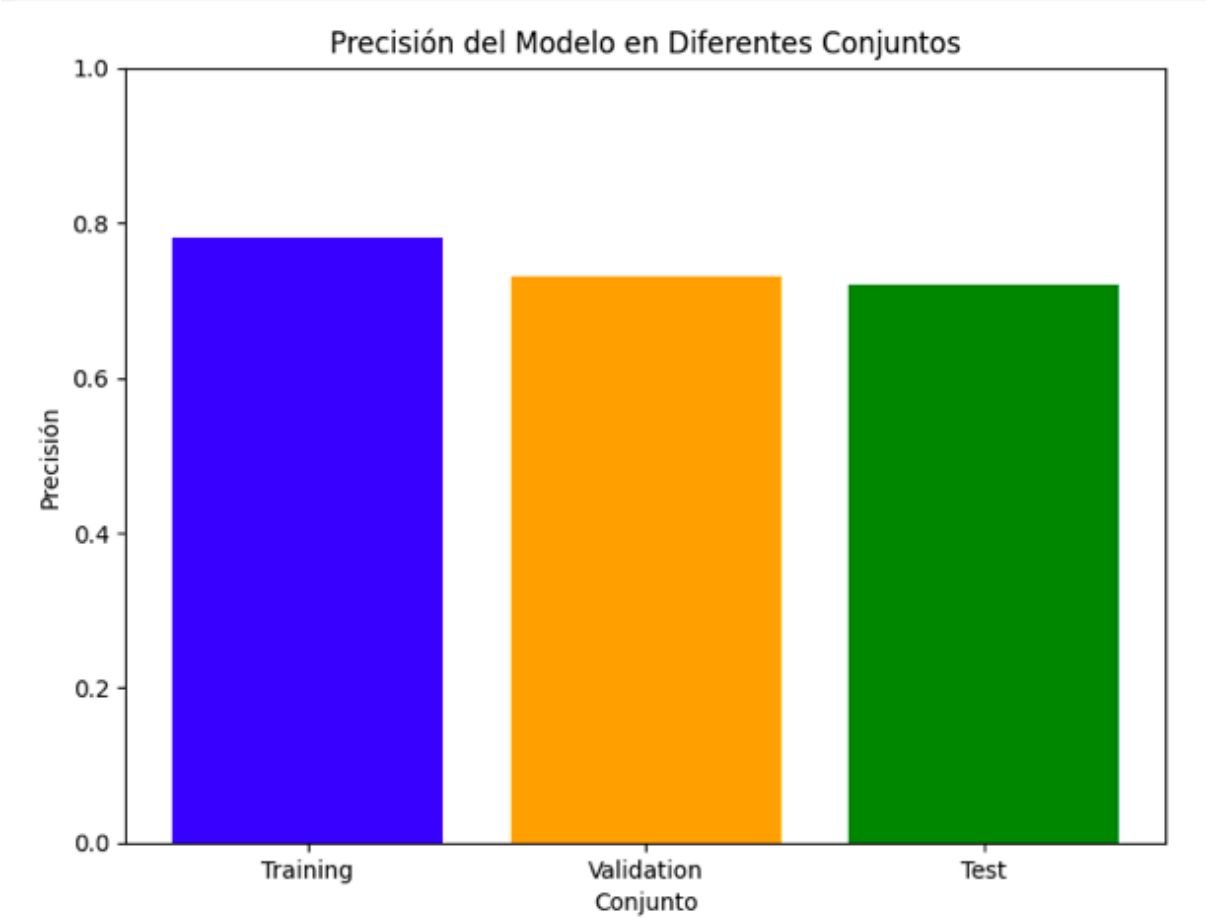
Diagnóstico y explicación del grado de bias

Entrene el modelo, evalué los tres conjuntos, visualicé y grafiqué las métricas para poder determinar el grado del sesgo .

```
Validation Accuracy: 0.8018867924528302
Validation Classification Report:
              precision    recall  f1-score   support

           B       0.78      0.94      0.86       200
           M       0.86      0.56      0.68       118

    accuracy                           0.80       318
   macro avg       0.82      0.75      0.77       318
weighted avg       0.81      0.80      0.79       318
```

```
Validation Accuracy: 0.7375
Validation Classification Report:
              precision    recall  f1-score   support

           B       0.73      0.92      0.81        49
           M       0.78      0.45      0.57        31

    accuracy                           0.74        80
   macro avg       0.75      0.68      0.69        80
weighted avg       0.75      0.74      0.72        80
```

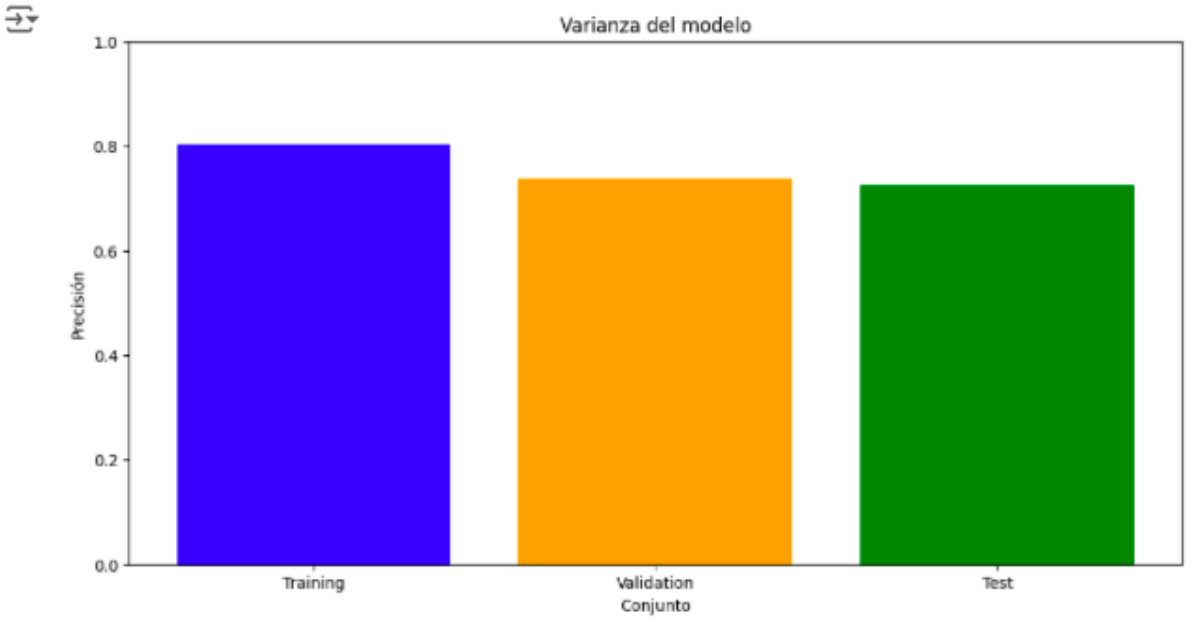
```
Test Accuracy: 0.7251461988304093
Test Classification Report:
              precision    recall  f1-score   support

           B       0.72      0.92      0.81       108
           M       0.74      0.40      0.52        63

    accuracy                           0.73       171
   macro avg       0.73      0.66      0.66       171
weighted avg       0.73      0.73      0.70       171
```



Para poder determinar el grado del sesgo tenemos que visualizar los score de ‘precision’ de los tres conjuntos. En los tres conjuntos tenemos rendimientos relativamente buenos y las métricas están bastante cerca unas de las otras, es decir, la precisión del conjunto de entrenamiento está bastante cerca de la precisión de los conjuntos de validación y de prueba, por lo tanto tenemos un sesgo relativamente bajo. Si hubiera gran diferencia entre los score de los conjuntos de validación y de prueba contra los de entrenamiento podríamos estar frente a un sesgo medio y si fuera demasiado diferencia frente a un overfitting, pero no es el caso. Después de este análisis podemos decir que el modelo generaliza bien, aunque si notamos que el modelo puede no ser el mejor ya que el accuray no es tan alto.

Diagnóstico y explicación del grado de varianza

En esta parte para poder determinar el grado de varianza me fije en la precisión de los conjuntos y en la diferencia de las barras de los gráficos, entonces calcule la diferencia y realice las gráficas.




```
Diferencia de Precisión entre Entrenamiento y Validación: 0.06
Diferencia de Precisión entre Entrenamiento y Prueba: 0.08
```

En este caso observamos una varianza baja, esto ya que la precisión de los conjuntos son similares, las diferencias entre sí son pequeñas, por lo tanto las barras del gráfico están bastante cerca unas de otras, y de igual forma la diferencia de precisión entre el conjunto de entrenamiento con los conjuntos de validación y test son pequeñas, por lo tanto presenciamos una varianza baja, esto en sí nos dice nuevamente que el modelo es generalmente bueno, hace un buen ajuste para los datos no vistos.

Diagnóstico y explicación del nivel de ajuste del modelo.

Para este análisis observe las gráficas y las métricas que ya obtuve previamente para la varianza y para el grado de bias. Observe nuevamente precisiones en los tres conjuntos son relativamente altas y similares, entonces no parece haber un sobreajuste, solamente los conjuntos de validación y prueba son

ligeramente menores que el de entrenamiento, entonces nuevamente se puede decir que el modelo generaliza bien y se está haciendo un buen ajuste en los datos no vistos.

Uso de técnicas de regularización o ajuste de parámetros para mejorar el modelo.

En este punto claramente notamos que el modelo no es el mejor ya que con un accuracy de 75% aproximadamente no podemos decir que es muy confiable, idealmente en un modelo buscamos más de 90% para poder tomarlo como válido. Como sabemos el algoritmo de KNN depende de la distancia entre los puntos de datos, entonces si los datos no están normalizados, dichas distancias pueden ser distorsionadas por características con rangos más amplios, por eso ajustamos las características de los datos para que tengan una escala similar, es decir, estandarizamos los datos (media 0 y desviación estándar de 1), también utilice la normalización min-max la cual escala las características para que los valores de estas estén entre 0 y 1, y finalmente implemente la normalización por escala de varianza la cual ajusta los datos para que tengan una varianza de 1.

```
Accuracy Estandarización: 0.9707602339181286
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.98       108
           1       0.97      0.95      0.96        63

    accuracy                           0.97       171
   macro avg       0.97      0.97      0.97       171
weighted avg       0.97      0.97      0.97       171
```

```
Accuracy normalizacion min-max: 0.935672514619883
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95       108
           1       0.96      0.86      0.91        63

    accuracy                           0.94       171
   macro avg       0.94      0.92      0.93       171
weighted avg       0.94      0.94      0.93       171
```

```
Accuracy normalización por escala de varianza: 0.9590643274853801
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.97       108
           1       0.94      0.95      0.94        63

    accuracy                           0.96       171
   macro avg       0.95      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```

Como podemos observar la mejor técnica es la estandarización, ya que obtenemos un accuracy de 97%, esto se debe a que los datos se están aproximando a una distribución normal, cosa que pudimos notar desde el gráfico que realizamos para observar la distribución de cuando dividimos los conjuntos, entonces de cierta manera ya desde ahí podemos notar que tipo de técnica utilizar o probar, realmente las otras dos las implementamos para poder hacer una comparación pero en este caso observamos fácilmente cual es el mejor, además de que aumentó el rendimiento del modelo considerablemente.