Guatemala 14 de marzo de 204
Arturo Argueta -21527
Astrid Glauser - 21299

Minería de datos

HT5

---

3. Analice los resultados del modelo de regresión. ¿Qué tan bien le fue prediciendo?

```
Accuracy of model: 0.684931506849315%
```

El resultado es pésimo , no es del 1%, pues el modelo requiere más preparación de los datos y naive bayes es para clasificación no para predecir una variable de respuesta directamente

4. Compare los resultados con el modelo de regresión lineal y el árbol de regresión que hizo en las hojas pasadas. ¿Cuál funcionó mejor?

el que funcionó mejor es la regresión lineal con una precisión del 80% (R^2) , pues es el más óptimo para predicciones de valores de una variable respuesta, mientras que naive bayes y árbol es para clasificación, sin embargo el arbol obtuvo un accuracy de 72%

(5 y 6). Haga un modelo de clasificación, use la variable categórica que hizo con el precio de las casas (barata, media y cara) como variable respuesta. Utilice los modelos con el conjunto de prueba y determine la eficiencia del algoritmo para predecir y clasificar.

Output del performance:

```
Accuracy: 0.6917808219178082
Confusion Matrix:
[[64  0  6]
 [ 0 76  5]
 [35 44 62]]
Classification Report:
              precision    recall  f1-score   support

       Caras       0.65      0.91      0.76        70
  Económicas       0.63      0.94      0.76        81
 Intermedias       0.85      0.44      0.58       141

    accuracy                           0.69       292
   macro avg       0.71      0.76      0.70       292
weighted avg       0.74      0.69      0.67       292
```

Comparación y Análisis

- El modelo de clasificación tiene una precisión general más alta (69.18%) en comparación con el modelo adaptado para regresión (56.16%).
- En el modelo de clasificación, las clases "Caras" y "Económicas" tienen altos valores de recall, lo que indica que el modelo es bueno para identificar estas categorías. Sin embargo, la clase "Intermedias" tiene un recall bajo, lo que sugiere que el modelo tiene dificultades para identificar correctamente esta categoría.
- En el modelo adaptado para regresión, la clase "Caras" tiene un recall muy alto (0.94), pero la clase "Intermedias" tiene valores bajos tanto en precisión como en recall, lo que indica una dificultad significativa para clasificar correctamente esta categoría.
- Ambos modelos tienen un rendimiento variable en diferentes categorías, pero el modelo de clasificación parece ser más equilibrado en términos de precisión y recall para las tres categorías.
- El modelo de clasificación es más eficiente para predecir y clasificar, especialmente para las categorías "Caras" y "Económicas". El modelo adaptado para regresión tiene un rendimiento más bajo, haciendo énfasis en la categoría "Intermedias".

7. Haga un análisis de la eficiencia del modelo de clasificación usando una matriz de confusión.Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.

Clase "Caras"

- **Verdaderos Positivos (TP) = 64**: El modelo clasificó correctamente 64 casas como "Caras".
- **Falsos Negativos (FN) = 6:** El modelo no identificó 6 casas que eran "Caras".
- **Falsos Positivos (FP) = 0**: El modelo no clasificó incorrectamente ninguna casa como "Cara".

tiene una alta precisión y un buen recall, lo que indica que el modelo es muy confiable al identificar casas "Caras" y rara vez las confunde con otras categorías.

Clase "Económicas"

- **Verdaderos Positivos (TP) = 76**: El modelo clasificó correctamente 76 casas como "Económicas".
- **Falsos Negativos (FN) = 5**: El modelo no identificó 5 casas que eran "Económicas".
- **Falsos Positivos (FP) = 0**: El modelo no clasificó incorrectamente ninguna casa como "Económica".

Tiene una alta precisión y un buen recall, lo que indica que el modelo es muy efectivo al clasificar casas en esta categoría.

Clase "Intermedias"

- **Verdaderos Positivos (TP) = 62:** El modelo clasificó correctamente 62 casas como "Intermedias".
- **Falsos Negativos (FN) = 79:** El modelo no identificó 79 casas que eran "Intermedias".
- **Falsos Positivos (FP) = 79:** El modelo clasificó incorrectamente 79 casas como "Intermedias".

Presenta el mayor desafío para el modelo, con una precisión más baja y un recall moderado. Esto indica una considerable confusión entre esta categoría y las otras.

Importancia de los Errores

Impacto de los Falsos Positivos (FP)

Los FP en la clase "Intermedias" pueden llevar a una sobreestimación de la calidad o el valor de las casas. Esto podría resultar en decisiones de inversión equivocadas.

Impacto de los Falsos Negativos (FN)

Los FN en la clase "Intermedias" pueden tener el efecto contrario, subestimando el valor de las casas que realmente pertenecen a esta categoría. Esto podría resultar en oportunidades perdidas o en la subvaloración de propiedades.

8. Analice el modelo. ¿Cree que pueda estar sobre ajustado?

```
Accuracy en el conjunto de entrenamiento: 0.7243150684931506
Accuracy en el conjunto de prueba: 0.6917808219178082
```

La diferencia entre la precisión en el conjunto de entrenamiento y la precisión en el conjunto de prueba es relativamente pequeña ,aproximadamente 3.25%. Esto indica que el modelo está generalizando bien a los datos no vistos y no muestra signos evidentes de sobreajuste

9. Haga un modelo usando validación cruzada, compare los resultados de este con los del modelo anterior. ¿Cuál funcionó mejor?

```
Puntajes de validación cruzada: [0.6369863  0.69178082 0.73630137 0.70205479 0.71232877]
Promedio de puntajes de validación cruzada: 0.695890410958904
Accuracy del modelo anterior en el conjunto de prueba: 0.6917808219178082
```

El promedio de los puntajes de validación cruzada es ligeramente superior a la precisión del modelo anterior en el conjunto de prueba. La diferencia es pequeña, sin embargo, el modelo evaluado con validación cruzada funcionó ligeramente mejor

10. Tanto para los modelos de regresión como de clasificación,pruebe con varios valores de los hiperparámetros, use el mejor modelo del tunneo, ¿Mejoraron los modelos? Explique

El modelo de clasificación no mejora, sino empeora perdiendo toda su accuracy, siendo de 0.006 , es decir menor al 1%. Y el de regresión aumenta a 1.02%, pero igualmente es insignificante sus predicciones, pues no es viable. Esto se debe a que al buscar los parámetros más influyentes se pierde precisión o arruina completamente el modelo.

11. Compare la eficiencia del algoritmo con el resultado obtenido con el árbol de decisión (el de clasificación) y el modelo de random forest que hizo en la hoja pasada. ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?

Tiempo de entrenamiento: 0.011999130249023438 segundos

El modelo de Random Forest para regresión tiene un R^2 de 0.8899, lo que indica un buen ajuste al conjunto de datos. En comparación, el modelo Naive Bayes para clasificación tiene una precisión del 69.18%