Universidad del Valle de Guatemala
Minería de Datos Sección 10
Lynette García

Hoja de Trabajo 5

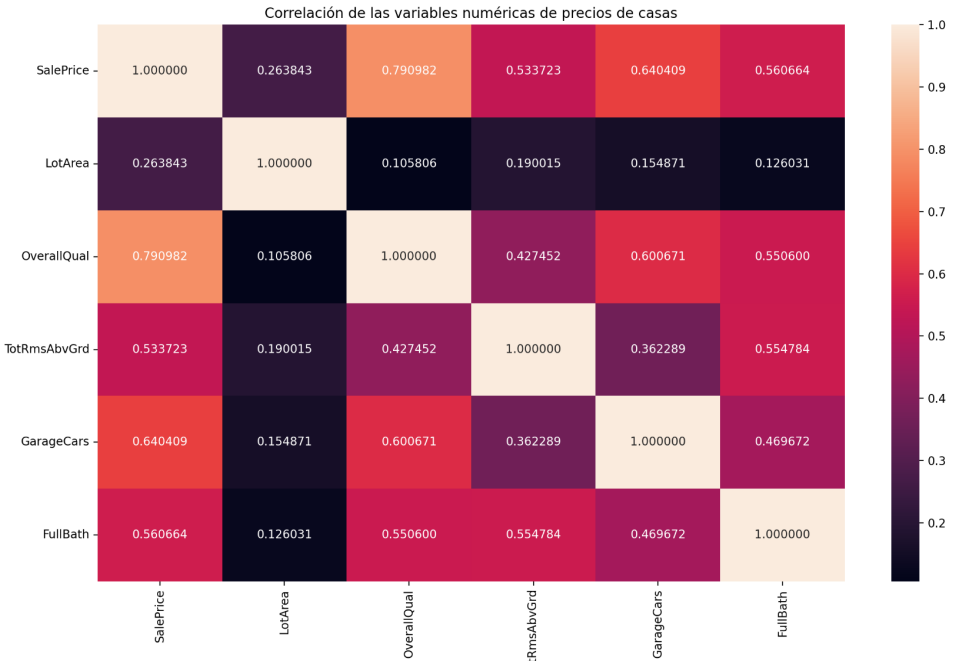
Grupo 7

16 de marzo del 2022

Sofía Rueda	Carné: 19099
Oliver De León	Carné: 19270
Martín España	Carné: 19258

Análisis de los modelos generados

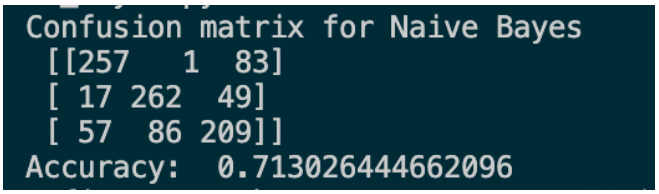
	SalePrice	LotArea	OverallQual	TotRmsAbvGrd	GarageCars	FullBath
SalePrice	1.000000	0.263843	0.790982	0.533723	0.640409	0.560664
LotArea	0.263843	1.000000	0.105806	0.190015	0.154871	0.126031
OverallQual	0.790982	0.105806	1.000000	0.427452	0.600671	0.550600
TotRmsAbvGrd	0.533723	0.190015	0.427452	1.000000	0.362289	0.554784
GarageCars	0.640409	0.154871	0.600671	0.362289	1.000000	0.469672
FullBath	0.560664	0.126031	0.550600	0.554784	0.469672	1.000000



Análisis de las variables a incluir en el modelo, pruebas de normalidad, correlación, etc.

Las variables utilizadas fueron ‘OverallQual’, ‘LotArea’, ‘TotalBathrooms’, ‘TotalBdrAbvGrd’

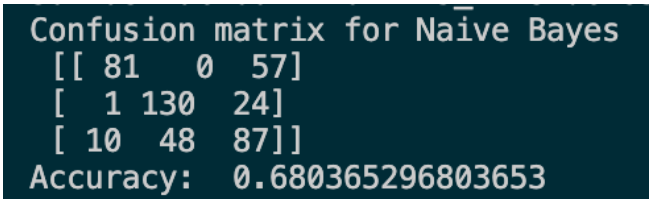
Aplicación de los modelos al conjunto de prueba



Modelo de naive bayes con el conjunto de entrenamiento

Como se puede apreciar en la imagen anterior, la precisión de las predicciones fue de 0.71 (71%) por lo que se puede considerar que fue satisfactoria.

Matriz de confusión de cada modelo y explicación de resultados

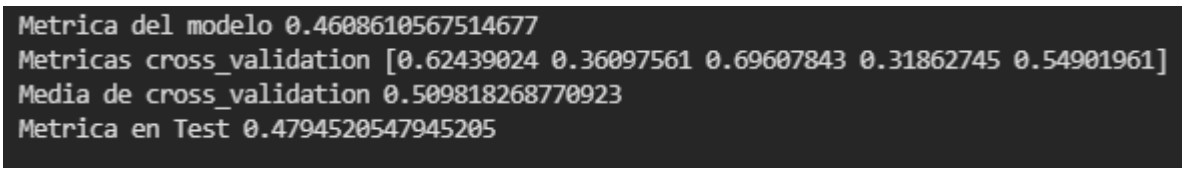


Modelo de naive bayes con el conjunto de entrenamiento

Como se puede apreciar en la imagen anterior, la precisión de las predicciones fue de 0.68 (68%) por lo que se puede considerar que fue satisfactoria.

Modelo de Comparación Cruzada y su comparación de precisión

La técnica de Validación Cruzada nos ayudará a medir el comportamiento de los modelos creados, ayudándonos a encontrar un mejor modelo rápidamente. Al reproducir la validación cruzada conforme al conjunto entrenado, encontramos los siguientes resultados frutos del testeo.

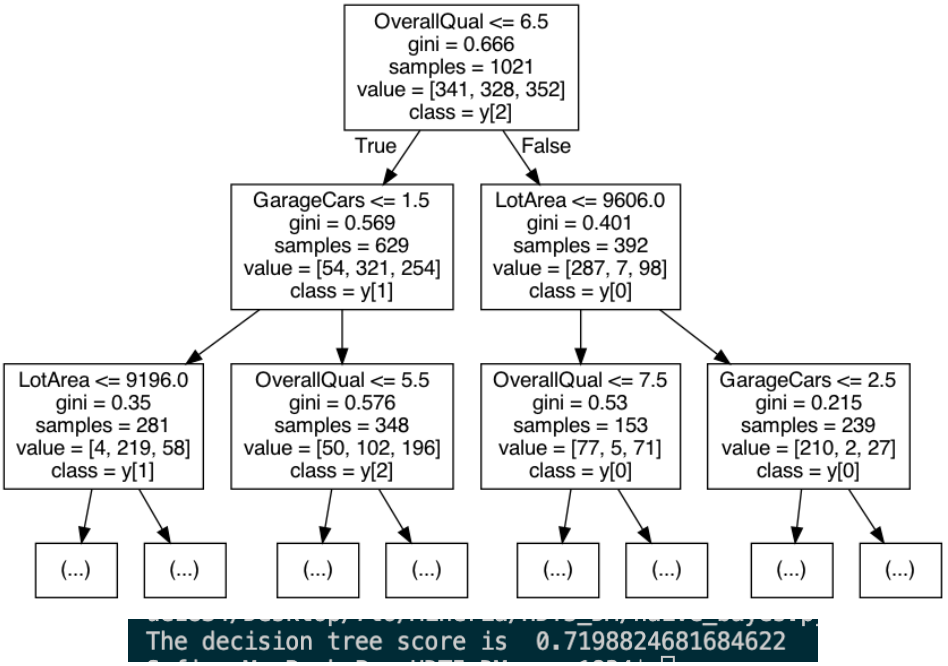


Brindándonos una medición del **48%** en cuanto a la precisión del modelo efectuado.

Comparación del método de naive bayes con el árbol de clasificación

```
Confusion matrix for Naive Bayes
 [[ 81   0  57]
 [  1 130  24]
 [ 10  48  87]]
Accuracy:  0.680365296803653
```

Matriz de confusión de naive bayes del conjunto de entrenamiento



Árbol de decisión (clasificación) del conjunto de entrenamiento

A simple vista es posible observar que el método de naive bayes es más eficiente y sencillo de implementar que el árbol de clasificación. Esto principalmente se debe a la asunción de que las variables son independientes entre sí, lo cual hace el procedimiento mucho más rápido que si se consideran las probabilidades de todas las variables dada una correlación entre ellas. No obstante, se puede observar que la precisión del algoritmo es cercana al 70%, lo cual es menor que con el árbol de decisión, que si bien parece ser más tardado de implementar y de ejecutar es más preciso con las predicciones.

Ahora, mucha precisión en las predicciones tampoco es buena como se ha mencionado a lo largo del curso, ya que podría tratarse de un Overfitting durante la fase de entrenamiento, y esto ocasiona que el modelo no sea muy efectivo al enfrentarse a datos del mundo real. Respecto al resultado de ambos métodos, es posible apreciar que el método del árbol de clasificación es mucho más fácil de visualizar gráficamente, pues cuenta con un grafo que muestra todas las decisiones tomadas durante el algoritmo. Sin embargo, el método de Naive Bayes tiene la ventaja de que, si bien el resultado no es muy atractivo o cómodo visualmente hablando, la información está resumida en la matriz de confusión de manera que se pueden apreciar todos los casos que ocurrieron durante la predicción (falsos negativos, verdaderos negativos, verdaderos positivos, falsos positivos).

Por otro lado, un posible problema que tiene el método de Naive Bayes es que mientras más variables se deseen utilizar, mayor será el "error" de la asunción mencionada anteriormente, por lo que el resultado probablemente se alejará más y más de los valores esperados. Aunque si se busca eficiencia para conjuntos de datos moderados este método probablemente sea el mejor a seguir por su facilidad de uso (Varghese, 2018).

Otra diferencia significativa en ambos métodos es que el árbol de decisión es discriminatorio mientras que Naive Bayes es generativo lo que significa que los árboles de clasificación son más flexibles en general y aceptan una mayor variedad de información. Por ello, para decidir cuál de los dos métodos es más conveniente utilizar, es necesario conocer la naturaleza de los datos. Para esto, se debe considerar si las variables a utilizar son independientes, en cuyo caso la opción más coherente sería utilizar Naive Bayes. En el caso contrario lo mejor sería usar el árbol de clasificación debido a la asunción mencionada.

En conclusión, ambos algoritmos son eficientes y fáciles de implementar, por lo que la mejor manera de discernir cuál de los dos es más conveniente utilizar dependerá de la información y de la naturaleza de los datos. Esto debería ser posible de determinar desde el análisis exploratorio, por lo que si se llegara a implementar alguno de los métodos y no resulta ser tan efectivo, probablemente deba revisarse que el análisis exploratorio se haya hecho de la mejor manera posible, considerando la correlación entre las variables, así como su dependencia (Antonidis, 2021).

Link de Repositorio en GitHub

https://github.com/srue1834/HDT5_DM

Referencias

- Antonidis, P. (2021). Decision Tree vs Naive Bayes Classifier. Extraído de: https://www.baeldung.com/cs/decision-tree-vs-naive-bayes
- Varghese, D. (2018). Comparative Study on Classic Machine Learning Algorithms. Extraído de: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=Decision%20tree%20vs%20naive%20Bayes,the%20accuracy%20for%20a%20toss.