Informe Proyecto 2 entrega 4

Juan Luis Solórzano (carnet: 201598) Micaela Yataz (carnet: 18960)

2025-01-20

git: https://github.com/JusSolo/Mineria_Proyecto2.git

1. Elabore un modelo de regresión usando K nearest Neighbors (KNN), el conjunto de entrenamiento y la variable respuesta SalesPrice. Prediga con el modelo y explique los resultados a los que llega. Asegúrese que los conjuntos de entrenamiento y prueba sean los mismos de las entregas anteriores para que los modelos sean comparables.

```
y<- datos$SalePrice
set.seed(123)
trainI<- createDataPartition(y, p=0.7, list=FALSE)

train<-datosC[trainI, ]
test<-datosC[-trainI, ]

## Modelo:
train<-train[complete.cases(train),]
k_vecinos <- round(sqrt(nrow(datosC)),0)
parametros <- expand.grid(k = k_vecinos)

modelo_knn1 <- train(
  SalePrice~.,
  data=train,
  method = "knn",
  preProcess= c("center","scale","knnImpute"),
  tuneGrid = parametros
)
predModelo1 <- predict(modelo_knn1, newdata=test)
```

2. Analice los resultados del modelo de regresión usando KNN. ¿Qué tan bien le fue prediciendo? Utilice las métricas correctas.

```
pred_test <- predict(modelo_knn1, newdata = test)

metrics <- postResample(pred_test, test$SalePrice)
print(metrics)
```

```
##         RMSE     Rsquared          MAE
## 4.056178e+04 7.835675e-01 2.336609e+04
```

El $R^2 = 0.784$ es aceptable, El MAE es del orden de los $20,000 lo que para el precio de una casa parece aceptable.

3. Compare los resultados con el modelo de regresión lineal, el mejor modelo de árbol de regresión y de naive bayes que hizo en las entregas pasadas. ¿Cuál funcionó mejor?

```
##              Modelo      RMSE       MAE        R2         r
## RMSE...1        KNN  40561.78  23366.09 0.7835675 0.8851935
## RMSE...2     Lineal  36792.87  21871.94 0.7772167 0.8815989
## RMSE...3      Árbol  46340.50  30161.08 0.6470656 0.8044039
## RMSE...4 NaiveBayes 196887.52 180897.46 0.6168137 0.7853749
```

Podemos notar que para todas las medidas de error excepto el MAE el mejor modelo es el KNN, el segundo mejor es el modelo lineal, seguido por el Arbol y el peor es el Naïve Bayes.

4. Haga un modelo de clasificación, use la variable categórica que hizo con el precio de las casas (barata, media y cara) como variable respuesta.

5. Utilice los modelos con el conjunto de prueba y determine la eficiencia del algoritmo para predecir y clasificar.

6. 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.

7. Analice el modelo. ¿Cree que pueda estar sobreajustado?

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

9. Tanto para los modelos de regresión como de clasificación, pruebe con varios valores de los hiperparámetros ¿Qué parámetros pueden tunearse en un KNN?, use el mejor modelo del tuneo, ¿Mejoraron los resultados usando el mejor modelo ahora? Explique

10. Compare la eficiencia del algoritmo con el resultado obtenido con el árbol de decisión (el de clasificación), el modelo de random forest y el de naive bayes que hizo en las entregas pasadas. ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?