**Trabajo Práctico N°1 - Grupo XX**

**Análisis Exploratorio**

Realizar una breve descripción del dataset: cantidad de registros y columnas, etc.

Comentar los features más destacables: tipo de dato, qué representa y por qué se destacan. Listar hipótesis o supuestos que tomaron.

**Preprocesamiento de Datos**

Detallar las tareas más importantes que realizaron sobre el dataset, les dejamos algunas preguntas cómo guía:

* ¿Se eliminaron columnas (Nombre de la columna y motivo de eliminación?
* ¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?
* ¿Generaron nuevos features?
* ¿Encontraron valores atípicos?¿Cuáles?¿Qué técnicas utilizaron y qué decisiones tomaron?
* ¿Qué columnas tenían datos faltantes?

¿En qué proporción? ¿Qué se hizo con estos registros?

**Visualizaciones**

Mostrar dos gráficos realizados: dispersión entre variables, histogramas, heatmaps, etc. que sean descriptivos del problema. Seleccionar aquellos que permitan entender cómo se distribuyen los datos, cómo se relacionan la variables etc.

Comentar brevemente qué se está visualizando en cada caso y por qué los eligieron.

**Clustering**

Mencionar la tendencia al clustering y la cantidad apropiada de grupos que se deben formar. Mostrar cómo llegaron a esa conclusión. Mostrar un gráfico sobre un mapa de CABA con los avisos coloreados según el grupo al que pertenecen. Describir en función de qué características se formaron dichos grupos.

**Clasificación**

Mencionar cuál es la alternativa seleccionada para construir la variable “tipo\_precio” justificando su elección. Mostrar en un mapa de CABA los avisos coloreados por tipo\_precio ¿Qué diferencias o similitudes encuentran con el agrupamiento realizado por K-Means con 3 grupos?

* **Construcción del modelo**

Arbol de Decision

* Optimizamos hiperparámetros, y utilizamos randomcv.best\_params\_ y randomcv.best\_score\_ con F1 Score
* Usamos K-fold Cross Validation y utilizamos folds=15
* De Metrica usamos make\_scorer(f1\_score, average='micro')
* imagen del árbol generado



Random Forest

* hiperparámetros usamos random\_cv\_random\_forest\_version.best\_params\_
* usamos folds = 5
* make\_scorer(f1\_score)

Modelo Regresion logistica

* ¿Optimizaron hiperparámetros? ¿Cuáles?
* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.

**Nota**: Para cada modelo mencionar si realizaron nuevas transformaciones sobre los datos (encoding, normalización, etc)

* **Cuadro de Resultados**

Realizar un cuadro de resultados comparando los modelos que entrenaron (entre ellos debe figurar cuál es el que seleccionaron como mejor predictor).

Medidas de rendimiento en el conjunto de TEST:

* F1
* Precision
* Recall
* Accuracy
* XXX: si seleccionaron alguna métrica adicional…

Confeccionar el siguiente cuadro con esta información:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | F1-Test | Precision Test | Recall Test | Accuracy Test | XXX |
| Arbol de Decision |  |  |  |  |  |
| Random Forest |  |  |  |  |  |
| … |  |  |  |  |  |

En cada caso ¿Cómo resultó la performance respecto al set de entrenamiento?

**Nota: indicar brevemente en qué consiste cada modelo de la tabla, por ejemplo**

Arbol de decision: {'max\_depth': 7, 'criterion': 'entropy', 'ccp\_alpha': 0.02}

Random Forest: {'criterion': 'entropy', 'min\_samples\_leaf': 1, 'min\_samples\_split': 16, 'n\_estimators': 50}

* **Elección del modelo**

En base a los resultados obtenidos responder las siguientes preguntas

¿Qué modelo elegirían para clasificar el tipo de precio de las propiedades?¿ Por qué?

**Regresión**

Mencionar en cada caso si realizaron nuevas transformaciones sobre los datos (encoding, normalización, etc) completando los ítems a y b:

* **Construcción del modelo**

KNN

* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

XGBoost

* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo a elección

* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.
* **Cuadro de Resultados**

Realizar un cuadro de resultados comparando los modelos que entrenaron (entre ellos debe figurar cuál es el que seleccionaron como mejor predictor).

Medidas de rendimiento en el conjunto de TEST:

* MSE
* RMSE
* XXX: si seleccionaron alguna métrica adicional…

Confeccionar el siguiente cuadro con esta información:

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | MSE | RMSE | XXX |
| KNN |  |  |  |
| XGBoost |  |  |  |
| … |  |  |  |

En cada caso ¿Cómo resultó la performance respecto al set de entrenamiento?

**Nota: indicar brevemente en qué consiste cada modelo de la tabla**

* **Elección del modelo**

En base a los resultados obtenidos responder las siguientes preguntas:

¿Qué modelo elegirían para clasificar el tipo de precio de las propiedades?¿ Por qué?

**Conclusiones Finales**

Realizar las conclusiones correspondientes al trabajo realizado en su totalidad, destacando principalmente los aspectos que consideren más relevantes. Mencionar también conclusiones producto de la experimentación (que puede o no haber sido volcada en la entrega).

Comentar brevemente qué otras opciones hubiesen explorado y quedaron fuera del alcance de este trabajo: qué experimentos hubiesen querido realizar o qué técnicas hubiesen querido utilizar y no lo hicieron (por tiempo, complejidad, falta de teórica/práctica etc).

**Tiempo dedicado**

|  |  |  |
| --- | --- | --- |
| Integrante | Tarea | Prom. Hs Semana |
| Flavian Ferré | Clasificacion  Regresión  KNN,XGBoost  armado de reporte | 9 |
| Alan Richmond | Clasificacion  Entrenamiento y Prediccion  armado de reporte | 5 |
| Alan Mejia | armado de reporte |  |