Proyecto 2. Entrega 4.
K Nearest Neighbors (KNN)

INTRODUCCIÓN:

InmoValor S.A. es una empresa innovadora del sector inmobiliario que ha apostado por la transformación digital para ofrecer valoraciones precisas y objetivas de propiedades. Ante un mercado dinámico y competitivo, la compañía ha adoptado técnicas avanzadas de análisis y modelos de regresión para estimar el valor de inmuebles basándose en un amplio conjunto de datos que recopila información detallada de viviendas. Este dataset incluye variables clave como ubicación, tamaño, calidad constructiva y otros factores determinantes, lo que permite desarrollar modelos predictivos capaces de reflejar con mayor exactitud las condiciones del mercado.

La empresa ha decidido incorporar un equipo de analistas de datos con la finalidad de trabajar con el conjunto de datos "House Prices: Advanced Regression Techniques" para desarrollar modelos predictivos que proyecten de manera precisa el precio de las viviendas. Mediante el análisis de variables clave como la ubicación, el tamaño y la calidad de las propiedades, el equipo utilizará técnicas avanzadas de regresión para mejorar la estimación de valores inmobiliarios y facilitar la toma de decisiones estratégicas en el mercado de bienes raíces.

Vínculo del conjunto de datos a utilizar

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Resultados esperados en la Tercera Entrega:

Se espera la entrega de un informe detallado donde incluya:

- Los modelos de KNN tanto para predicción del precio de las casas como de clasificación usando la variable categórica de los precios que creó.
- Una comparación para estimar los precios de las viviendas donde determine qué algoritmo funcionó mejor (Regresión Lineal, Árboles de regresión, Random Forest, Bayes Ingenuo, KNN). Compare los resultados del mejor modelo de cada uno de los algoritmos. Recuerde que para establecer una comparación válida, es necesario comparar bajo las mismas condiciones. Mismo conjunto de datos de entrenamiento y de prueba.
- Un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos en la generación y aplicación de los modelos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado.
- Esta vez le han pedido un informe formal, que enlace los temas usando subtítulos explicativos que mantenga al lector en un hilo conductual de los temas que aborda la consultoría. No incluya los enunciados de las actividades de esta guía como subtítulos.

Notas:

- La consultoría es en grupo, por lo que solo se tendrán en cuenta los grupos conformados por más de un especialista.

- Cada individuo será evaluado de forma individual basado en sus aportes al trabajo grupal, por lo que deben versionar el código para poder revisar las contribuciones de cada uno.

ACTIVIDADES

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

EVALUACIÓN

Nota: Tiene que poderse comprobar su aporte al trabajo grupal a través de commits. Si no existen al menos 3 commits con su aporte significativo no va a tener nota de la hoja de trabajo. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(20 puntos)** Análisis de los modelos generados. Explique todos los razonamientos. Uso de las métricas correctas
- **(20 puntos)** Comparación del modelo de regresión con los modelos de regresión de las entregas pasadas. Elección del mejor modelo hasta el momento basado en métricas adecuadas.
- **(10 puntos).** Tuneo de parámetros. Se probó con varios valores para los hiperparámetros. Se explican los resultados.

- **(20 puntos)** Matriz de confusión de cada modelo de clasificación. Explicación de los resultados obtenidos. Selección del mejor modelo usando las métricas adecuadas. Análisis de otras métricas además del accuracy para los modelos generados.
- **(20 puntos)** Comparación del modelo de clasificación de KNN con el mejor modelo de clasificación de todas las entregas pasadas. Elección del mejor modelo de clasificación basado en las métricas adecuadas.
- **(10 puntos)** El informe de resultados está bien redactado, se establece una continuidad lógica en las explicaciones y el análisis de los modelos. Tiene un hilo conductor claro que no pierde al lector y no tiene las actividades de esta guía como subtítulos.

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con lo solicitado en las instrucciones
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó .rmd debe añadir el .html que se genera)
- Link del repositorio utilizado para la hoja de trabajo.

FECHAS DE ENTREGA

- **AVANCE:** Puntos del 1 al 5 de la sección de actividades: viernes 21 de marzo a las 23:59.
- **ENTREGA FINAL:** domingo 23 de marzo a las 23:59