Proyecto 2. Entrega 6.
Máquinas Vectoriales de Soporte (SVM)

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 Quinta Entrega:

Se espera la entrega de un informe detallado donde incluya:

- Los modelos de SVM (Support Vector Machine) para clasificación usando la variable categórica de los precios que creó.
- Los modelos de SVR (Support Vector Regressor) para predicción de los precios de las casas.
- Una comparación para estimar los precios de las viviendas donde determine qué algoritmo funcionó mejor (Árboles de regresión, Random Forest, Bayes Ingenuo, KNN, Regresión Lineal, SVR). 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.
- Una comparación para estimar la categoría de los precios de las viviendas donde determine qué algoritmo funcionó mejor (Árboles de decisión, Random Forest, Bayes Ingenuo, KNN, Regresión Logística, SVM). 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.
- Como la vez anterior, 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. Use los mismos conjuntos de entrenamiento y prueba de las hojas de trabajo pasadas para probar el algoritmo.
2. Explore los datos y explique las transformaciones que debe hacerle para generar un modelo de máquinas vectoriales de soporte.
3. Use como variable respuesta la variable categórica que especifica si la casa es barata, media o cara
4. Genere varios (más de 2) modelos de SVM con diferentes kernels y distintos valores en los parámetros c, gamma (circular) y d (en caso de que utilice el polinomial). Puede tunear el modelo de forma automática siempre que explique los resultados
5. Use los modelos para predecir el valor de la variable respuesta
6. Haga las matrices de confusión respectivas.
7. Analice si los modelos están sobreajustados o desajustados. ¿Qué puede hacer para manejar el sobreajuste o desajuste?
8. Compare los resultados obtenidos con los diferentes modelos que hizo en cuanto a efectividad, tiempo de procesamiento y equivocaciones (donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores).
9. Compare la eficiencia del mejor modelo de SVM con los resultados obtenidos en los algoritmos de las hojas de trabajo anteriores que usen la misma variable respuesta (árbol de decisión y random forest, naive bayes, KNN, regresión logística). ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?
10. Genere un buen modelo de regresión, use para esto la variable del precio de la casa directamente. Tunee el modelo.
11. Compare los resultados del modelo de regresión generado con los de hojas anteriores que utilicen la misma variable, como la de regresión lineal, el árbol de regresión, naive bayes, KNN.
12. Genere un informe de los resultados y las explicaciones.

EVALUACIÓN

Notas:

- 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.
- Debe entregar los avances durante el período de clase para poder tener derecho a la calificación de la hoja de trabajo.

- **(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 SVM 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 .r o .py con el código y hallazgos comentados
- Link de Google docs con las conclusiones y hallazgos encontrados. Puede usar también Jupyter Notebooks o rmd.
- Vínculo del repositorio usado para trabajar la hoja de trabajo.

FECHAS DE ENTREGA

- **AVANCE:** Puntos del 1 al 7 de la sección de actividades: jueves 24 de abril a las 17:20.
- **ENTREGA FINAL:** domingo, 27 de abril a las 23:59