Hoja de Trabajo 3.
Árboles de Decisión

INTRODUCCIÓN:

Kaggle

Kaggle es una comunidad en línea de científicos de datos, propiedad de Google LLC. Permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web, trabajar con otros científicos de datos e ingenieros de aprendizaje automático, y participar en competencias para resolver los desafíos de la ciencia de datos. Tuvo su inicio al ofrecer competencias de aprendizaje automático y ahora también ofrece una plataforma pública de datos, una mesa de trabajo basada en la nube para la ciencia de la información y educación en IA de formato corto. El 8 de marzo de 2017, Google anunció que estaban adquiriendo Kaggle.

Conjunto de datos a utilizar

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

Notas:

- La hoja de trabajo se realizará en parejas.
- Los grupos serán seleccionados por afinidad.
- La hoja no se calificará si no pertenece a ningún grupo de los creados en canvas para esta hoja.

INSTRUCCIONES

Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir los precios de las casas. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual.

ACTIVIDADES

1. Descargue los conjuntos de datos de la plataforma kaggle.
2. Haga un análisis exploratorio extenso de los datos. Explique bien todos los hallazgos. No ponga solo gráficas y código. Debe llegar a conclusiones interesantes para poder predecir. Explique el preprocesamiento que necesitó hacer.
3. Incluya un análisis de grupos en el análisis exploratorio. Explique las características de los grupos.

4. Dependiendo del análisis exploratorio elaborado cree una variable respuesta que le permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben tener un fundamento en la distribución de los datos de precios, y estar bien explicados.
5. Divida el set de datos preprocesado en dos conjuntos: Entrenamiento y prueba. Describa el criterio que usó para crear los conjuntos: número de filas de cada uno, estratificado o no, balanceado o no, etc. Si le proveen un conjunto de datos de prueba y tiene suficientes datos, tómelo como de validación, pero haga sus propios conjuntos de prueba.
6. Elabore el árbol de clasificación utilizando el conjunto de entrenamiento y la variable respuesta que creó en el punto 4. Explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.
7. Elabore el árbol de regresión para predecir el precio de las viviendas utilizando el conjunto de entrenamiento. Explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.
8. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar y predecir, en dependencia de las características de la variable respuesta.
9. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.
10. Analice el desempeño del árbol de regresión.
11. Repita los análisis usando random forest como algoritmo de predicción, explique sus resultados comparando ambos algoritmos.

EVALUACIÓN

- **(8 puntos)** Análisis exploratorio. Recuerde explicar los razonamientos.
- **(8 puntos)** Análisis de los grupos
- **(10 puntos)** Creación de la variable respuesta para árbol de clasificación. Explicación de los límites de las categorías.
- **(8 puntos)** Conjuntos de Entrenamiento y prueba. Descripción del método usado para crearlos
- **(12 puntos)** Árbol de Clasificación. Representación gráfica del modelo.
- **(12 puntos)** Árbol de Regresión. Representación gráfica del modelo.
- **(12 puntos)** Random Forest
- **(30 puntos)** Análisis de resultados de aplicación del algoritmo para predecir o clasificar sobre el conjunto de prueba. Comparación entre algoritmos.

MATERIAL A ENTREGAR

- Archivo .r o .py con el código y hallazgos comentados
- Archivo .pdf con las conclusiones y hallazgos encontrados. (Opcional, puede incluir comentarios en el archivo de código)