Práctica 6.5

Entrega: durante la clase

Descargue el conjunto de datos titanic.csv y lea la descripción del mismo en `https://www.kaggle.com/competitions/titanic/data`. Este conjunto de datos será usado en los siguientes puntos, los cuales deben ser resueltos usando scikit-learn.

1. Prepare los datos.
   - *a*) Cargue los datos usando Pandas.
   - *b*) Selecciones las siguientes características del dataset: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked.
   - *c*) Imprima las primeras filas del dataframe es inspeccione los datos. Qué variables son categóricas y cuáles son numéricas?
   - *d*) Verifique cuántos datos faltantes hay usando las funciones isnull and sum de Pandas
   - *e*) Utilice la función fillna de Pandas para manejar valores faltantes, use la mediana o la moda según aplique.
   - *f*) Convierta las columnas con valores categóricos a valores numéricos usando la función get_dummies de Pandas.
   - *g*) Cree dos dataframes X y y para los datos de entrada y la variable de clase (*Survived*)
2. Entrene un modelo de clasificación basado en árboles de decisión.
   - *a*) Haga una partición del conjunto de datos, usando muestreo estratificado, en 70 % para entrenamiento y 30 % para test.
   - *b*) Entrene el modelo
   - *c*) Aplique el modelo al conjunto de test
   - *d*) Mida el desempeño del modelo calculando exactitud, error de clasificación, precisión, recall, curva ROC y matriz de confusión.
3. Interprete el modelo obtenido:
   - *a*) Grafique el árbol obtenido
   - *b*) Cuál el el atributo más discriminante? Tiene sentido? De una explicación a partir del conocimiento del problema.
4. Comparación de modelos:
   - *a*) Usando el mismo conjunto de datos, entrene un modelo random forest.
   - *b*) Compare el modelo con el árbol de decisión.
   - *c*) Cuál de los dos modelos es mejor?

El taller debe enviarse como un Jupyter notebook a través del siguiente Dropbox file request, antes de la medianoche de la fecha límite. El archivo debe nombrarse como isi-practica6.5-unalusername.ipynb, donde unalusername es el nombre de usuario asignado por la universidad. No incluya archivos adicionales.