Universidad de Buenos Aires
Facultad De Ingeniería
Año 2021 - 2do Cuatrimestre

Organización de Datos (75.06)

TRABAJO PRÁCTICO Nº2
TEMA:
FECHA DE ENTREGA: 8/12/2021

INTEGRANTES:

DE LUCA ANDREA, Felipe	- #88888
FOPPIANO, Elián	- #88888

Índice

1. Preprocesamientos

Nombre del procesamiento	Descripción	Nombre de la función
Común	Drop de *llovieron_hamburguesas_hoy* Reemplazar *si* y *no* por 1 y 0 en variable target Drop de muestras con missings en variable target Drop de muestras con datos inválidos Corrección de typos en nombres de features Si el parámetro **fecha_to_int** es True, convierte *dia* en un entero con formato AAAAMMDD	**common**()
Día a mes	Convierte *dia* en el mes correspondiente a la fecha	**dia_a_mes**()
Viento trigonométrico	Convierte las features de dirección del viento en pares de features $\langle \sin\theta, \cos\theta \rangle$, donde θ es el ángulo de dicha dirección	**viento_trigonometrico**()
Barrios a comunas	Realiza un hashing de los valores de la feature **barrio**, donde el nombre de cada barrio se reemplaza por el nombre de su comuna (*Comuna 1*, *Comuna 2*, etc)	**barrios_a_comunas**()
Estandarización	Agrega un **StandardScaler()** a la **Pipeline**, que transforma los datos tal que su distribución tenga esperanza 0 y varianza 1 (deben ser todas las features numéricas)	**standarizer**()
Imputador simple	Agrega un **SimpleImputer()** a la **Pipeline**, que imputa los valores faltantes con el promedio	**simple_imputer**()
Imputador iterativo	Agrega un **IterativeImputer()** a la **Pipeline**, que imputa los valores faltantes a partir de regresiones iterativas	**iterative_imputer**()
Drop categóricas	Dropea las features categóricas del dataset (Si se ejecutó el preprocesamiento *Viento trigonometrico*, solo dropea la feature **barrio**)	**drop_categoricas**()
Hashing trick	Transforma una feature categórica en varias columnas con valores 0 o 1 a partir de un vector obtenido con una función de hash sobre la variable	**drop_categoricas**()
Drop poco importantes	Dado una lista de scores y un threshold, dropea las features cuyos scores (importancia de feature) sea menor al threshold	**drop_poco_importantes**()
Feature selection	Dado un modelo y un porcentaje de features a mantener, selecciona las features que maximicen el score del modelo dado mediante forward selection	**feature_selection**()

Cuadro 1.1: Preprocesamientos

2. Modelos

Nombre del modelo	Preprocesamiento	AUC ROC	Accuracy	Precision	Recall	F1 score
DecisionTreeClassifier	Común (**fecha_to_int=True**) Viento trigonométrico Barrios a comunas Día a mes Imputador iterativo	0.870	0.849	0.763	0.470	0.582
RandomForestClassifier	Común (**fecha_to_int=False**) Viento trigonométrico Día a mes Drop categóricas Estandarización Imputador iterativo	0.889	0.859	0.764	0.538	0.631
SVM	Común (**fecha_to_int=True**) Viento trigonométrico Estandarización Imputador iterativo Hashing trick (feature **barrio**)	0.882	0.857	0.794	0.491	0.607
NB (Naive Bayes)	Común (**fecha_to_int=True**) Viento trigonométrico Drop categóricas Imputador iterativo Feature selection (**n = 0.6**)	0.850	0.832	0.644	0.565	0.602
KNN	Común (**fecha_to_int=True**) Viento trigonométrico Estandarización Imputador iterativo Hashing trick (feature **barrio**)	0.873	0.838	0.805	0.365	0.503
Perceptrón	Común (**fecha_to_int=True**) Viento trigonométrico Barrios a comunas Drop poco importantes (todas las features que tuvieron coeficientes 0 en un perceptrón previo) Imputador iterativo Estandarización	0.833	0.820	0.770	0.282	0.413
AdaBoostClassifier	Común (**fecha_to_int=True**) Barrios a comunas Viento trigonométrico Imputador simple Estandarización	0.894	0.861	0.806	0.498	0.615
CascadingClassifier	Común (**fecha_to_int=True**) Viento trigonométrico Hashing Trick (feature **barrio**) Imputador iterativo Estandarización	0.895	0.862	0.778	0.538	0.636
NeuralNetwork	Común (**fecha_to_int=True**) Viento trigonométrico Hashing Trick (feature **barrio**) Imputador iterativo Estandarización	0.895	0.860	0.774	0.529	0.629

Cuadro 2.1: Métricas de los modelos