

Reto sobre aplicación de Python para respuesta a preguntas de analítica

Profesores:

Christian Urcuqui (ulcamilo@gmail.com)

Universidad Icesi



Objetivos de esta sesión

- 1. Usar Python para la exploración de un conjunto de datos.
- Responder a un conjunto de preguntas a partir de los datos
- Aplicar Python para el desarrollo de funciones, ciclos, asignación de variables y explorar la información de un dataframe.



Taller práctico



Contexto

Análisis de datos

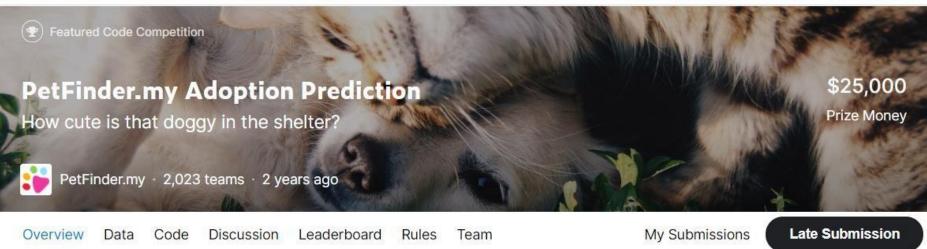
PetFinder.my Adoption Prediction How cute is that doggy in the shelter?

Millones de animales sufren en la calle o son sacrificados en refugios todos los días en todo el mundo. Si se pueden encontrar hogares para ellos, se pueden salvar muchas vidas preciosas — y se pueden crear más familias felices.

Esta competencia busca que se los participantes puedan proponer algoritmos que permitan predecir la capacidad de adopción de las mascotas, ¿con qué rapidez se adopta una mascota?

https://www.kaggle.com/c/petfinder-adoption-prediction





Overview

Description

Evaluation

Timeline

Prizes

Kernels-FAQ

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. If homes can be found for them, many precious lives can be saved — and more happy families created.

PetFinder.my has been Malaysia's leading animal welfare platform since 2008, with a database of more than 150,000 animals. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare.

Animal adoption rates are strongly correlated to the metadata associated with their online profiles, such as descriptive text and photo characteristics. As one example, PetFinder is currently experimenting with a simple AI tool called the Cuteness Meter. which ranks how cute a pet is based on qualities present in their photos.



PREGUNTAS DE ANÁLISIS

Vamos a utilizar solo el archivo train.csv Complejidad 1

- 1. ¿Cuantos registros y variables cuenta el dataset?
- 2. ¿Cómo están los valores de la variable Age?
 - 1. ¿Cuál es la distribución por tipo de animal
 - 2. Cuales son los estadísticos descriptivos para esta variable.
- 3. ¿Hay más datos de perros o de gatos?
- 4. ¿Cómo es la velocidad de adopción para perros y gatos?
- 5. Encuentre los nombres sin valores.
 - 1. ¿Cuántos para gatos y perros?
 - 2. ¿Eliminar o reemplazar los registros?, justifique
- 6. ¿la salud podría significar algo en la velocidad de 4/28/2021adopción?



PREGUNTAS DE ANÁLISIS

Vamos a utilizar solo el archivo train.csv

Complejidad 2

- 7. ¿Cuales son los cinco nombres más populares para los perros?
- Revise la relación entre la variable de edad y la velocidad de adopción.
- 9. ¿Existe alguna relación entre la longitud del nombre y la velocidad de adopción?



Complejidad 3

10. Ejecute las siguientes líneas de código e interprete

11. Ejecute las siguientes líneas e interprete

12. Ejecute e interprete los resultados de clasificación (precision and recall)

```
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)

from sklearn.metrics import classification_report

y_pred = clf.predict(X_test)

print(classification_report(y_test, y_pred))
```



Complejidad 3

13. Ejecute las siguientes líneas de código ¿Cuál es el propósito de mutual_info_classif?

```
discrete features = [True, False, False, False, False, False,
                         True, False, False, False,
                        True, True, False, False, False, False]
    from sklearn.feature selection import mutual info classif
    def make_mi_scores(X, y, discrete_features):
        mi scores = mutual info classif(X, y, discrete features= discrete features)
        mi scores = pd.Series(mi scores, name="MI Scores", index=X.columns)
  5
        mi scores = mi scores.sort values(ascending=False)
        return mi scores
  9 mi scores = make mi scores(X train, y train, discrete features)
10 mi scores[::3]
 1 def plot mi scores(scores):
        scores = scores.sort values(ascending=True)
        width = np.arange(len(scores))
        ticks = list(scores.index)
        plt.barh(width, scores)
        plt.yticks(width, ticks)
  6
        plt.title("Mutual Information Scores")
import matplotlib.pyplot as plt
```

```
import numpy as np

plt.figure(dpi=100, figsize=(8, 5))
4/28/2 plot_mi_scores(mi_scores)
```



Complejidad 3

14. Con base al anterior resultado seleccione cuatro variables, entrene y evalué un nuevo modelo con las mismas configuraciones y datos. Interprete el proceso y los nuevos resultados