Trabajo Practico:

1. Gregors:

Fecha entrega: 28/10/2023. (Esto fue una tarea que se dio hace dos semanas)

Las entregas se enumeran a continuación

a. EDA personal:

Enfocarse en las variables order_has_email y order_has_phone respecto a la conversión mensual y compararlo a las variables de horario segun zona geografica.

b. Diapositiva 5 Slides

Utiliza el Pyramid Principle de Barbra Minto para estructurar tus hallazgos en una presentación PowerPoint de 5 diapositiva

2. Resolución Olist.

Fecha entrega: 31/10/2023. (Esto fue una tarea que se dio hace dos semanas)

Crearse una repositorio personal en github. Completar el codigo dentro de la carpeta olist/ y completar las notebooks

- a. https://github.com/federicomoreno613/UCEMA_2023/blob/ma ster/part-3/2-ecommerce/ejercicio orders part1.ipynb
- b. https://github.com/federicomoreno613/UCEMA_2023/blob/master/part-3/2-ecommerce/ejercicio_orders_part2.ipynb

3. Regresion

https://drive.google.com/file/d/1PDYKuc0Ly7iBjG81voDPiVHmDY0o0Zmw/view?usp=sharing

Fecha entrega: 5/11/2023.

El dataset de startups contiene variables como gastos en investigación y desarrollo (I+D), costos administrativos, gastos en marketing, ubicación geográfica y beneficios netos.

Este tipo de datos es esencial para entender el rendimiento financiero y operativo de una startup, lo cual es crítico para la toma de decisiones tanto para los inversionistas como para la administración de la empresa.

- **a.** Análisis Exploratorio de Datos (EDA) ¿Qué conclusiones puede sacar con respecto a las distribuciones de las variables numéricas y las relaciones entre ellas?
- **b.** Modelo de Regresión Lineal para Predecir Beneficio Neto.
- **c.** Cual es la intuicion detrás de la formula de la regresion resultante.
- **d.** Cual es el error promedio? y el error elevado al cuadrado? ¿Que significa?
- e. Analice del Sector de la Industria en el Beneficio Neto.
- f. Cual es la relacion entre las series y el profit? Es importante hay alguna particularidad? (Hint)

4. Clasificación

Fecha entrega: 8/11/2023.

https://drive.google.com/file/d/1CAJMqRuYS_8-P1wO0lypkLvbBFf -bnlx/view?usp=sharing

Contenido_de_Nitrogeno: Proporción de contenido de nitrógeno en el suelo.

Contenido_de_Fosforo: Proporción de contenido de fósforo en el suelo. Contenido_de_Potasio: Proporción de contenido de potasio en el suelo.

Temperatura C: Temperatura en grados Celsius.

Humedad_Relativa: Humedad relativa en porcentaje.

Nivel_de_pH: Nivel de pH del suelo.

Precipitacion_mm: Cantidad de precipitación en milímetros.

Tipo_de_Cultivo: Tipo de cultivo recomendado en función de los parámetros anteriores.

El uso de Machine Learning en la agricultura es vital para optimizar rendimientos, mejorar la sostenibilidad y minimizar riesgos. Esto es particularmente relevante en Argentina, donde la agricultura es un pilar económico. Saber qué tipo de cultivo es más adecuado para ciertas condiciones del suelo y del clima puede tener un impacto significativo en la productividad y la sostenibilidad.

a. Análisis Exploratorio de Datos (EDA)

Para realizar un EDA exhaustivo del dataset, considere las siguientes preguntas:

- ¿Cómo se distribuyen los valores en cada una de las variables?
- ¿Existen correlaciones entre las diferentes variables del suelo y los tipos de cultivos?
- ¿Cómo varía la recomendación de cultivo según las condiciones del suelo y el clima?
- ¿Existen valores atípicos o faltantes que deban ser tratados?

b. Machine Learning

En esta sección, implementaremos dos modelos de Machine Learning para predecir el Tipo_de_Cultivo basándonos en las variables del suelo y del clima. Utilizaremos una Regresión Logística y un Árbol de Decisión.

Consigna para la Implementación del Código: Utilice el siguiente código de Python como base para implementar el modelo de Árbol de Decisión:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
import seaborn as sns

# Crear el modelo
decision_tree_model = DecisionTreeClassifier(criterion="entropy", random_state=2,
max_depth=5)

# Validación cruzada
score = cross_val_score(decision_tree_model, features, target, cv=5)
print('Puntuación de validación cruzada:', score)

# Precisión en entrenamiento
dt_train_accuracy = decision_tree_model.score(x_train, y_train)
print("Precisión en entrenamiento =", dt_train_accuracy)

# Precisión en pruebas
dt_test_accuracy = decision_tree_model.score(x_test, y_test)
print("Precisión en pruebas =", dt_test_accuracy)

# Matriz de confusión
y_pred = decision_tree_model.predict(x_test)
y_true = y_test
from sklearn.metrics import confusion_matrix
cm_dt = confusion_matrix(y_true, y_pred)

# Visualización de la matriz de confusión
f, ax = plt.subplots(figsize=(15, 10))
sns.heatmap(cm_dt, annot=True, linewidth=0.5, fmt=".0f", cmap='viridis', ax=ax)
plt.xlabel("Predicho")
plt.ylabel("Real")
plt.title('Matriz de Confusión')
plt.title('Matriz de Confusión')
plt.title('Matriz de Confusión')
```

Interprete que significa la diagonal y los desvios de la diagonal. Cual cultivo se confunde mas el algoritmo?