Spring 2024:CS5720

Neural Networks and Deep Learning - ICP-3

Gonaboyina Vijay Vardhan (700755141)

Github link: https://github.com/Vijayvardhan02/Neural-networks-and-deep-learning-ICP3

Video:https://drive.google.com/file/d/1WBKrMcoWoAQv5B3C23jd7_iEa0Op7zFv/view?usp=drive_link

1. Data Manipulation

a. Read the provided CSV file 'data.csv'.

b.https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing

c. Show the basic statistical description about the data.

d. Check if the data has null values. i. Replace the null values with the mean

e. Select at least two columns and aggregate the data using: min, max, count, mean.

f. Filter the dataframe to select the rows with calories values between 500 and 1000.

g. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.

h. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse".

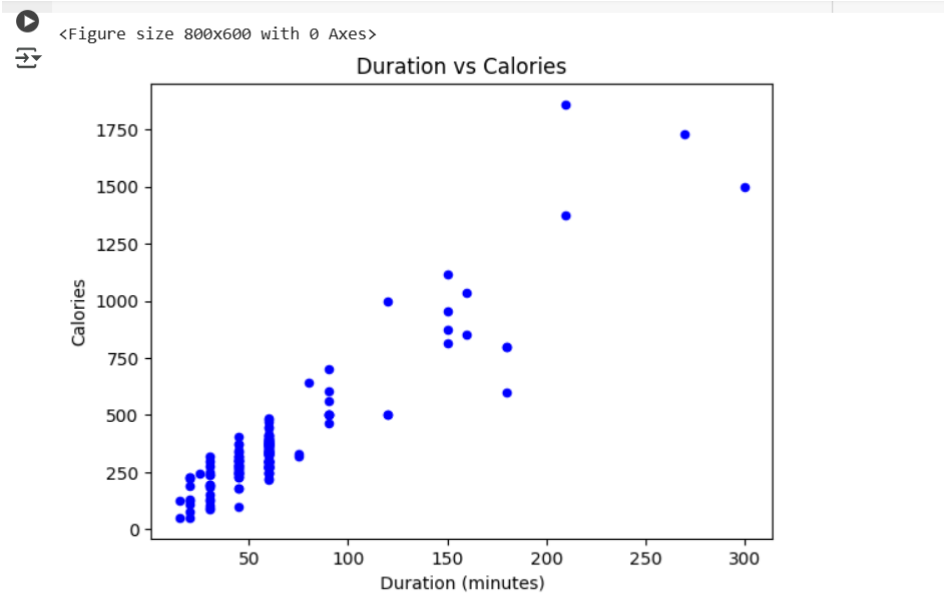
i. Delete the “Maxpulse” column from the main df dataframe j. Convert the datatype of Calories column to int datatype. k. Using pandas create a scatter plot for the two columns (Duration and Calories).

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import io

try:
    df = pd.read_csv('data.csv')
except FileNotFoundError:
    from google.colab import files
    uploaded = files.upload()
    df = pd.read_csv(io.BytesIO(uploaded['data.csv']))


df.fillna(df.mean(), inplace=True)

agg_df = df[['Duration', 'Calories']].agg(['min', 'max', 'count', 'mean'])

filtered_df_1 = df[(df['Calories'] >= 500) & (df['Calories'] <= 1000)]

filtered_df_2 = df[(df['Calories'] > 500) & (df['Pulse'] < 100)]

df.drop(columns=['Maxpulse'], inplace=True)

df['Calories'] = df['Calories'].astype(int)

plt.figure(figsize=(8, 6))
df.plot(kind='scatter', x='Duration', y='Calories', title='Duration vs Calories', color='blue')
plt.xlabel('Duration (minutes)')
plt.ylabel('Calories')
plt.show()
```

OUTPUT:

```
<Figure size 800x600 with 0 Axes>
```



2. Linear Regression

a) Import the given “Salary_Data.csv”

b) Split the data in train_test partitions, such that 1/3 of the data is reserved as test subset.

c) Train and predict the model.

d) Calculate the mean_squared error

e) Visualize both train and test data using scatter plot.

Code:

```
[ ]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_squared_error
     import io

     try:
         df = pd.read_csv('Salary_Data.csv')
     except FileNotFoundError:
         from google.colab import files
         uploaded = files.upload()
         filename = list(uploaded.keys())[0]
         df = pd.read_csv(io.BytesIO(uploaded[filename]))

     X = df[['YearsExperience']]
     y = df['Salary']

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

     model = LinearRegression()
     model.fit(X_train, y_train)

     y_pred = model.predict(X_test)

     mse = mean_squared_error(y_test, y_pred)
     print(f'Mean Squared Error: {mse}')

     plt.scatter(X_train, y_train, color='blue', label='Train Data')
     plt.scatter(X_test, y_test, color='red', label='Test Data')
```

```
plt.plot(X_test, y_pred, color='green', label='Regression Line')

plt.title('Linear Regression: Train and Test Data')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.legend()
plt.show()
```

OUTPUT:

Choose Files Salary_Data.csv

- **Salary_Data.csv**(text/csv) - 454 bytes, last modified: 1/28/2025 - 100% done

```
Saving Salary_Data.csv to Salary_Data.csv
Mean Squared Error: 35301898.887134895
```