Regression and Classification Analysis Using Python

150544 محمود عبدالله فؤاد طرابلسي

143397 احمد محمد عبدالجواد قمحية

161524 مؤمن محمد علي العمري

1. Introduction

This project demonstrates the application of machine learning techniques on two real-world datasets. The primary objectives include:

- **Regression Analysis:** Predicting numerical outcomes using the Corona Virus dataset.
- **Classification Task:** Categorizing data using the Asthma dataset.

The project incorporates data preprocessing, pipelines, and model evaluation to ensure robust and efficient workflows. Both single test dataset evaluation and cross-validation methods are used to compare model performance.

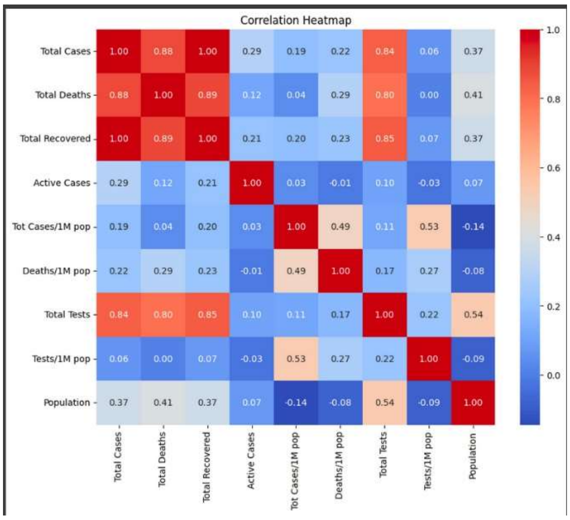
---

2. Dataset Description

We selected two datasets for this project:

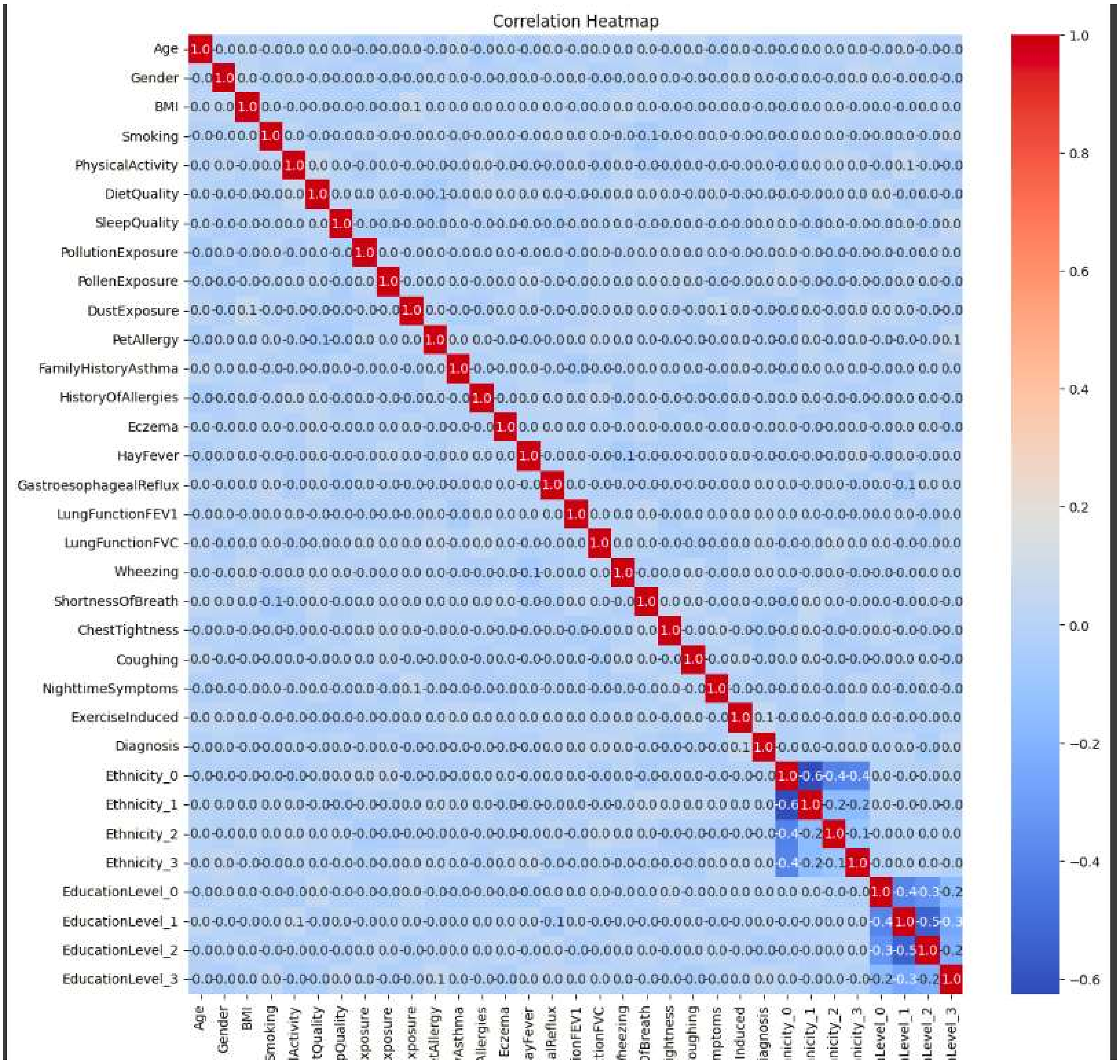
2.1 Regression Dataset

- **Dataset Name:** Corona Virus Dataset
- **Description:** This dataset includes features related to COVID-19 statistics, and the target variable is the total number of deaths.
- **Summary:**
    - Number of Rows: 232
    - Number of Columns: 13
    - Missing Values: 879
    - Data Types: Numerical



2.2 Classification Dataset

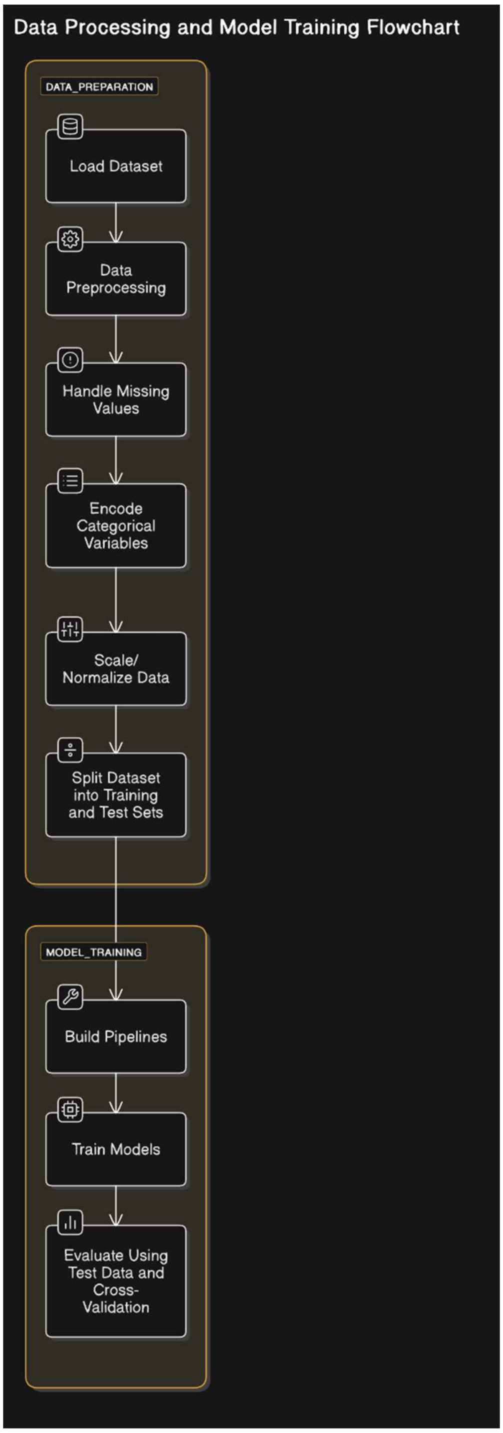
- **Dataset Name:** Asthma Dataset
- **Description:** This dataset contains features such as BMI, sleep quality, lung function metrics, and the target variable {Diagnosis} indicates whether a patient has asthma.
- **Summary:**
    - Number of Rows: 2392
    - Number of Columns: 29
    - Missing Values: 0
    - Data Types: Categorical, Numerical



3. Methodology

3.1 Flowchart

A flowchart of our approach is outlined below:


Data Processing and Model Training Flowchart
DATA_PREPARATION
Load Dataset
Data Preprocessing
Handle Missing Values
Encode Categorical Variables
Scale/ Normalize Data
Split Dataset into Training and Test Sets
MODEL_TRAINING
Build Pipelines
Train Models
Evaluate Using Test Data and Cross-Validation

3.2 Preprocessing

We applied the following preprocessing steps:

- **Handling Missing Values:** in the corona data set we dropped the columns that had more than 40% missing values.
- **Encoding Categorical Variables:** Used one-hot encoding for classification datasets.
- **Scaling/Normalization:** Applied Min-Max Scaling for regression and classification datasets to ensure model performance.

3.3 Models

We applied these models to each dataset:

- **Regression Models:**
    1. Linear Regression
    2. Decision Tree Regressor
    3. Random Forest Regressor
    4. SVR (support vector machine)
    5. K-nearest Neighbors
- **Classification Models:**
    1. Logistic Regression
    2. Support Vector Machine (SVC)
    3. K-Nearest Neighbors classifier
    4. Decision tree classifier
    5. Random forest classifier
    6. Naïve Bayes

3.4 Parameters

Default parameters were used initially.

4. Results

4.1 Regression Results

Model	Test R^2 Score	Mean Cross-Validation R^2
Linear Regression	1.0	1.0
Decision Tree Regressor	0.493	-0.5507
Random Forest Regressor	0.606	-11.25

4.2 Classification Results

Model	Test Accuracy	Mean Cross-Validation Accuracy
Logistic Regression	0.958	0.948
Support Vector Machine	0.9582	0.948
K-Nearest Neighbors	0.956	0.946

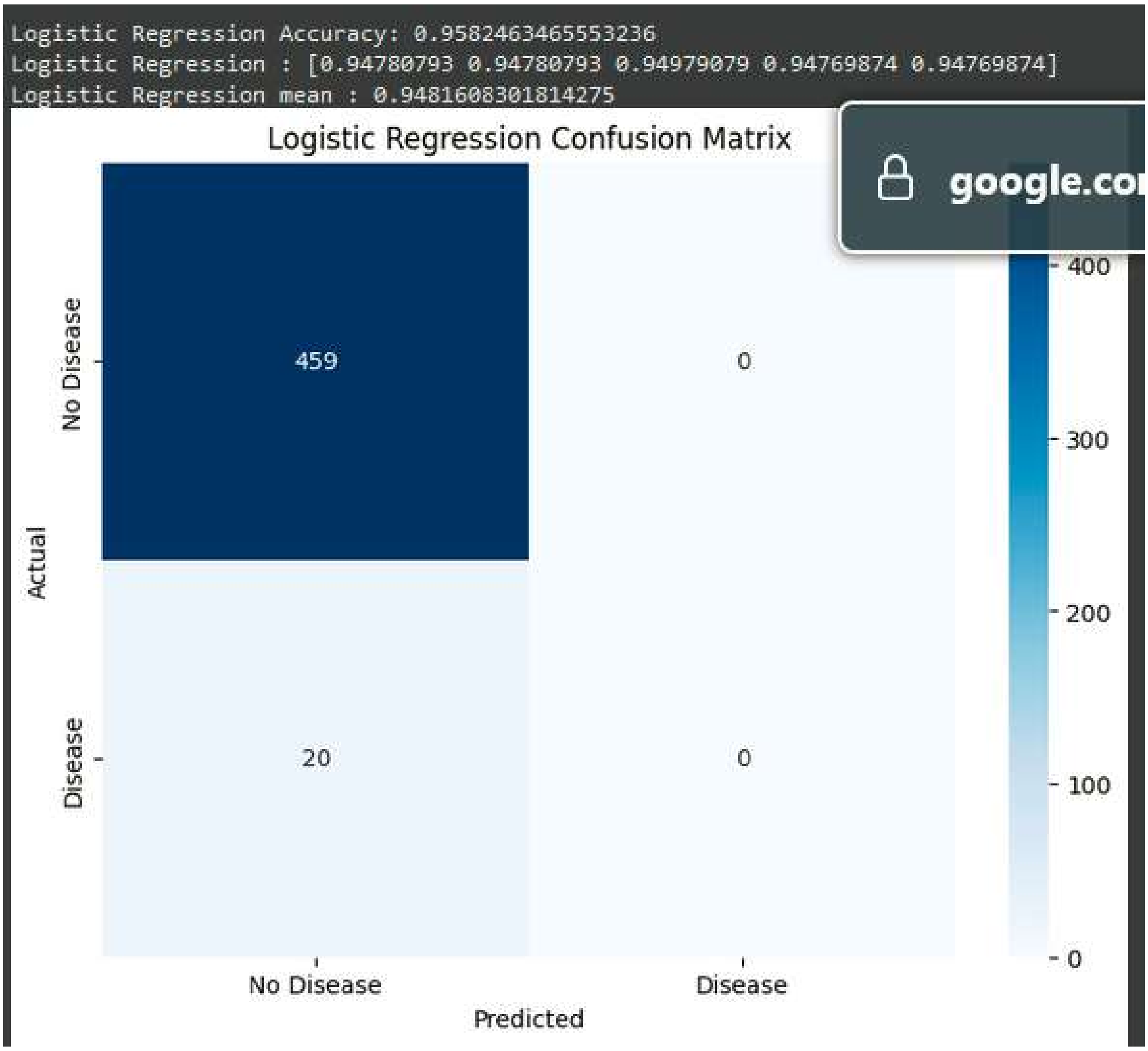
4.3 Discussion of Results

Regression:

- **Linear Regression performed excellently, with both test and cross-validation R^2 scores at 1.0, indicating a perfect fit for the data.**
- **Decision Tree Regressor and Random Forest Regressor showed poor performance, especially in cross-validation. The negative values in cross-validation R^2 for both models (e.g., Decision Tree: -0.5507 and Random Forest: -11.25) suggest that these models overfitted to the training data. In simple terms, they captured noise or irrelevant patterns in the data rather than the true underlying relationship.**
  - **The negative R^2 occurs when the model performs worse than a simple mean-based prediction. It indicates that the model is not generalizing well to unseen data, resulting in poor performance on cross-validation folds.**
  - **This issue could also be due to the high number of missing values in the Corona Virus dataset, which may lead to instability in tree-based models without proper handling of missing data or feature engineering.**

Classification:

- **Logistic Regression and SVM performed well with consistent results, while KNN showed slightly lower accuracy but was still strong.**
- **Random Forest and Decision Tree Classifiers performed reasonably well, but their results were slightly lower, suggesting that simpler models could perform just as well for this task.**



```
Logistic Regression Accuracy: 0.9582463465553236
Logistic Regression : [0.94780793 0.94780793 0.94979079 0.94769874 0.94769874]
Logistic Regression mean : 0.9481608301814275
```

Cross-Validation:

Cross-validation highlighted the overfitting issue in regression models, where models like Decision Tree and Random Forest did well on the training data but failed to generalize, leading to negative R^2 scores.

5. Conclusion

The negative cross-validation scores in the regression task reflect overfitting, where complex models failed to generalize to unseen data. Linear Regression performed the best, while tree-based models (e.g., Decision Tree and Random Forest) need further tuning or preprocessing to improve generalization.

Code Co-Lab Link

Asthma model

https://colab.research.google.com/drive/1h6CCehtIw2vktLDyfbT1HhkeFzjBAFad?usp=sharing

coronavirus model

https://colab.research.google.com/drive/1FqMxzHnKtOzw7lITS47re8bq3bf4OnyQ?usp=sharing