Introduction

In this project, we will be participating in Kaggle competitions, which involve challenging
machine learning tasks organized by Kaggle and other institutions. Our focus will be on two
datasets: the "Diabetes Prediction Dataset" and the "Corona Virus Latest Data 2023." Our goal
is to develop accurate predictive models and compete with other data scientists on the platform.
We will explore different approaches, experiment with techniques, and compare the performance of our
models to showcase our machine learning skills.

Data explanation

1- Corona Virus Dataset:
The "Corona Virus" dataset contains information about the coronavirus outbreak. It includes data like the number of confirmed cases, deaths, recoveries, testing rates, and vaccination progress. This dataset helps researchers understand how the virus spreads and its impact on different regions. By analyzing the dataset, data scientists can develop models to predict future outbreak trends and provide insights for decision-making.

2- Diabetes Prediction Dataset:
The "Diabetes Prediction" dataset focuses on predicting diabetes occurrence. It includes information like age, BMI, blood pressure, glucose level, and family history. By analyzing this dataset, data scientists can develop models to predict the likelihood of an individual having diabetes based on their characteristics. This helps healthcare professionals identify people at risk and implement preventive measures or personalized interventions for better management of diabetes.

Methodology and steps

Decision Tree Classifier:

1- Data Loading: The datasets, "Corona Virus" and "Diabetes Prediction," were loaded into the project.

2- Data Exploration: Basic exploratory analysis was performed to understand the datasets structure, features, and patterns.

3- Preprocessing: Preprocessing steps, such as handling missing values, feature encoding, and feature scaling, were applied to prepare the data for model training.

4- Feature Selection: Relevant features were selected for each dataset, and the target variable was separated from the feature set.

5- Model Training and Evaluation: The Decision Tree Classifier was used to build a predictive model for each dataset. This classifier
is a popular machine learning algorithm that learns decision rules from the training data and creates a tree-like model. The model
was trained on the training data and evaluated using accuracy score, classification report, and confusion matrix.

Methodology and steps

Logistic Regression:

1- Data Loading: The datasets, "Corona Virus" and "Diabetes Prediction," were loaded into the project.

2- Data Exploration: Basic exploratory analysis was performed to understand the datasets' structure, features, and patterns.

3- Preprocessing: Preprocessing steps, such as handling missing values, feature encoding, and feature scaling, were applied to prepare the data for model training.

4- Feature Selection: Relevant features were selected for each dataset, and the target variable was separated from the feature set.

5- Model Training and Evaluation: The Logistic Regression model was used to build a predictive model for each dataset. Logistic Regression
is a statistical model used for binary classification. The model was trained on the training data and evaluated using accuracy score,
classification report, and confusion matrix.