CS362 ARTIFICIAL INTELLIGENCE

Second Semester 2022-2023

**Final Project**

**Team members:**

**Member (Mohammed Ahmad )**

**Member ( Khaled Samara )**

**Member ( Ibrahem baniissa )**

**introduction:**

This project focuses on participating in Kaggle competitions, which are challenging machine learning tasks organized by Kaggle and other institutes. Our team will work with two datasets: the "Diabetes Prediction Dataset" and the "Coronavirus Latest Data 2023."

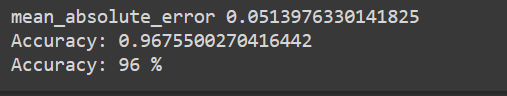
The "Diabetes Prediction Dataset" allows us to develop a model that predicts the likelihood of an individual having diabetes using features like age, BMI, and blood pressure. By leveraging advanced machine learning techniques, we aim to build an accurate predictive model for early diabetes detection.

The "Coronavirus Latest Data 2023" provides up-to-date information on infection rates, fatalities, and vaccination progress related to COVID-19. Analyzing this dataset will help us uncover patterns and trends to better understand the virus's spread and contribute to effective mitigation strategies.

Using Python programming language and essential libraries, our goal is to compare different approaches and models for accurate predictions. Participating in these competitions allows us to contribute to data science while honing our skills in machine learning and predictive modeling.

diabetes\_prediction\_dataset -> Classification

**Random Forests** (Kaggle)



**Confusion matrix**

A [confusion matrix](https://en.wikipedia.org/wiki/Confusion_matrix) is a way to express how many of a classifier’s predictions were correct, and when incorrect, where the classifier got confused (hence the name!). In the confusion matrices below, the rows represent the true labels and the columns represent predicted labels. Values on the diagonal represent the number (or percent, in a normalized confusion matrix) of times where the predicted label matches the true label. Values in the other cells represent instances where the classifier mislabeled an observation; the column tells us what the classifier predicted, and the row tells us what the right label was. This is a convenient way to spot areas where the model may need a little extra training.

A screenshot of numbers

Description automatically generated with low confidence

# Classification report

To get even more insight into model performance, we should examine other metrics like precision, recall, and F1 score.

**Precision** is the number of correctly identified members of a class divided by all the times the model predicted that class. In the case of Aspens, the precision score would be the number of correctly identified Aspens divided by the total number of times the classifier predicted “Aspen,” rightly or wrongly.

**Recall** is the number of members of a class that the classifier identified correctly divided by the total number of members in that class. For Aspens, this would be the number of actual Aspens that the classifier correctly identified as such.

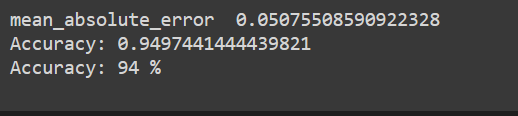
**F1 score** is a little less intuitive because it combines precision and recall into one metric. If precision and recall are both high, F1 will be high, too. If they are both low, F1 will be low. If one is high and the other low, F1 will be low. F1 is a quick way to tell whether the classifier is actually good at identifying members of a class, or if it is finding shortcuts (e.g., just identifying everything as a member of a large class).

A screenshot of a screen

Description automatically generated with low confidence

# Confusion matrix and Classification report (how to evaluate Random Forest model) are Cross-posted from Prof. [Jenny R. Kreiger](https://jrkreiger.net/) blog’s (Senior Data Scientist)

Decision Tree (Kaggle)



**Confusion matrix**

A screenshot of a computer

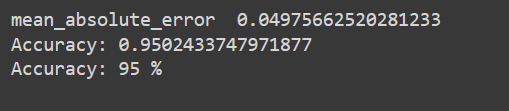
Description automatically generated with low confidence

# Classification report

A screenshot of a computer screen

Description automatically generated with low confidence

KNN (Kaggle)



**Confusion matrix**

A picture containing text, font, screenshot, number

Description automatically generated

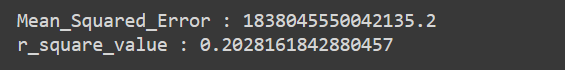
# Classification report

A picture containing text, screenshot, font, number

Description automatically generated

**coronavirus data set 🡪 Regression**

**linear Regression** (Kaggle)



**Decision Tree Regression**

****

**Methodology:**

For this project, we followed a systematic approach to tackle the machine learning tasks in the Kaggle competitions. Here is a brief overview of the methodology used:

**Diabetes Prediction - Classification Models:**

Data Collection: We obtained the "Diabetes Prediction Dataset" from Kaggle, which provides information on age, BMI, blood pressure, and other relevant features.

Data Preprocessing: We handled missing values, performed feature engineering, and split the data into training and testing sets.

Model Training and Evaluation: We employed Random Forests, Decision Trees, and K-Nearest Neighbors algorithms for classification. Each model was trained, and their performance was evaluated using metrics like accuracy, precision, recall, and F1-score.

Model Comparison: We compared the performance of the classification models and selected the best-performing one based on evaluation metrics.

Results and Visualization: We presented the findings using tables and charts to showcase the performance of the chosen model and provide insights into diabetes prediction.

**Coronavirus Analysis - Regression Models:**

Data Collection: We gathered the "Coronavirus Latest Data 2023" dataset, which contains up-to-date information on infection rates, fatalities, vaccination progress, and other relevant variables.

Data Preprocessing: We handled missing values, performed necessary transformations, and split the data into training and testing sets.

Model Training and Evaluation: We utilized Linear Regression, Random Forests, and Decision Trees algorithms for regression analysis. Each model was trained on the COVID-19 dataset, and their performance was evaluated using metrics like mean squared error and R-squared.

Model Comparison: We compared the performance of the regression models and identified the most effective one based on evaluation metrics.

Results and Visualization: We presented the findings using visualizations such as charts and tables to showcase the performance and insights gained from the selected regression model.