**TOPIC:** Diabetic prediction using multiple machine learning algorithms

**Dataset link:** [https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download](https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download%20)

* The project objective is to develop accurate predictive models for diabetic prognosis using various machine-learning algorithms.
* The analysis will involve 3 primary data science techniques:

1. classification
2. regression
3. clustering

* these techniques will leverage the medical and demographic features available in the data set.
* For **classification, we use** logistic regression or gradient boosting can categorize patients as diabetic or nondiabetic based on their medical and demographic information.

Logistic regression is chosen due to its effectiveness in binary classification problems and its ability to provide probabilities for class membership.

* **Regression analysis** will involve linear regression to predict continuous outcomes such as blood glucose levels and heart rate. Linear regression is suitable for this task due to its simplicity and interpretability, allowing for the determination of the relationship between predictor variables (e.g., age, BMI, HbA1c levels) and the outcome variable. The model’s performance will be assessed using metrics such as the mean squared error (MSE) to ensure accuracy.
* **Clustering** will be implemented using k-means clustering to identify patterns within the data. This method is chosen for its efficiency in handling large datasets and its capability to group patients with similar characteristics. By clustering patients, insights into distinct subgroups with varying risk profiles can be obtained, potentially informing targeted interventions.

Overall this project aims to utilize the power of multiple machine learning techniques to enhance the predictive accuracy of diabetic prognosis and provide valuable insights for healthcare applications.

**Logistic regression**: **Logistic regression** is a**supervised machine learning algorithm**used for **classification tasks** where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm that analyzes the relationship between two data factors.

Logistic regression is used for binary [classification](https://www.geeksforgeeks.org/getting-started-with-classification/) where we use [sigmoid function](https://www.geeksforgeeks.org/derivative-of-the-sigmoid-function/), that takes input as independent variables and produces a probability value between 0 and 1.

**linear regression**: Linear regression is a type of [supervised machine learning](https://www.geeksforgeeks.org/supervised-machine-learning/) algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

**Linear regression** is also a type of machine-learning algorithm more specifically a **supervised machine-learning algorithm** that learns from the labeled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

**mean squared error (MSE):** MSE is a way to quantify the accuracy of a model’s predictions. MSE is sensitive to outliers as large errors contribute significantly to the overall score.

is an evaluation metric that calculates the average of the squared differences between the actual and predicted values for all the data points. The difference is squared to ensure that negative and positive differences don’t cancel each other out.

𝑀𝑆𝐸=1𝑛∑𝑖=1𝑛(𝑦𝑖–𝑦𝑖^)2*MSE*=*n*1​∑*i*=1*n*​(*yi*​–*yi*​​)2

Here,

* n is the number of data points.
* yi is the actual or observed value for the ith data point.
* 𝑦𝑖^*yi*​​ is the predicted value for the ith data point.

**k-means clustering**: [K-Means Clustering](https://www.geeksforgeeks.org/k-means-clustering-introduction/) is an[Unsupervised Machine Learning](https://www.geeksforgeeks.org/ml-types-learning-part-2/) algorithm, which groups the unlabeled dataset into different clusters.

K-Means divides objects into clusters that group similar objects in one cluster and dissimilar objects into another.