Project idea and plan:

**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 **Diabetes prediction dataset** is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.

This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans.

Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

The Dataset can be utilized to construct machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details.

Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset. EHRs are digital versions of patient health records that contain information about their medical history, diagnosis, treatment, and outcomes. The data in EHRs is collected and stored by healthcare providers, such as hospitals and clinics, as part of their routine clinical practice

**COLLECTION METHODOLOGY**

The collection methodology for the diabetes prediction dataset involves gathering medical and demographic data from patients who have been diagnosed with or are at risk of developing diabetes. The data is typically collected through surveys, medical records, and laboratory tests.

The dataset can also be used for research purposes to identify potential risk factors for diabetes and to develop effective prevention and treatment strategies.

* The project objective is to develop accurate predictive models for diabetic prognosis using various machine-learning algorithms.

***Research Question***

* Which medical alongside demographic characteristics are the most crucial predictors of diabetes?
* How precisely can the following classifiers classify the respective patients as diabetic or the non-diabetic?
* What is the connection among the demographic or medical characteristics along with levels of blood glucose as predicted by the particular regressions?
* What distinct patient clusters may be recognised utilising the clustering, and also how do the diabetes risk profiles contradict?

***Project Objectives***

* To recognize crucial medical alongside demographic predictors of diabetes.
* To establish and develop the regressions model for precise diabetes classification.
* To develop regressions for predicting the blood glucose levels according to demographic along with medical characteristics.
* To implement clustering for recognizing distinct patient groups in varying diabetes risk profiles along with evaluating the characteristics.

***Summary of project and background***

Within this particular research, the purpose is for implementing precise predictive approaches for diabetic prediction utilizing an adequate dataset obtained from Kaggle. The specific dataset consists of 100,000 records having medical along with demographic data, for example, gender, age, hypertension, BMI, smoking history, heart disease, HbA1c level, along with blood glucose level. Different machine learning strategies, involving regression, classification, along with clustering, will be used for examining these elements as well as predict the diabetes risk (Tigga and Garg, 2020). The specific background of this particular project features the rising prevalence of diabetes overall and also the significance of the early diagnosis along with intervention. By utilizing machine learning approaches, critical predictors of the diabetes may be recognized, upgrading the comprehension of the risk factors and supporting the improvement of preventive techniques (Hasan *et al.* 2020). Classifiers can be utilized for the classification for evaluating if the patients are diabetic or non-diabetic. Regression may be utilized for predicting constant results like blood glucose levels, giving details into the connection among different features and also diabetes severity (Sumathy *et al.* 2022). The specific clustering can recognize distinct patient groups with comparative qualities, offering a more profound comprehension of various risk profiles.

***Reference List***

Tigga, N.P. and Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, pp.706-716. <https://www.sciencedirect.com/science/article/pii/S1877050920308024/pdf?md5=cc07853955b872e0f1553e48498a67d3&pid=1-s2.0-S1877050920308024-main.pdf>

Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516-76531. <https://ieeexplore.ieee.org/iel7/6287639/6514899/09076634.pdf>

Sumathy, B., Chakrabarty, A., Gupta, S., Hishan, S.S., Raj, B., Gulati, K. and Dhiman, G., 2022. Prediction of diabetic retinopathy using health records with machine learning classifiers and data science. International Journal of Reliable and Quality E-Healthcare (IJRQEH), 11(2), pp.1-16. <https://www.igi-global.com/viewtitle.aspx?titleid=299959>

* 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 val

[Diabetes prediction (Gradient Boosting) (kaggle.com)](https://www.kaggle.com/code/antoninmoreno/diabetes-prediction-gradient-boosting)

**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 is a** [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.

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

Support vector regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value.

SVR can use both linear and non-linear kernels. A linear kernel is a simple dot product between two input vectors, while a non-linear kernel is a more complex function that can capture more intricate patterns in the data. The choice of kernel depends on the data’s characteristics and the task’s complexity

**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.

## Random Forest

The [Random forest](https://www.geeksforgeeks.org/videos/random-forest-algorithm-in-machine-learning/) or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees. Random Forests are particularly well-suited for handling large and complex datasets, dealing with high-dimensional feature spaces, and providing insights into feature importance. This algorithm’s ability to maintain high predictive accuracy while minimizing overfitting makes it a popular choice across various domains, including finance, healthcare, and image analysis, among others.

## Random Forest Classifier

The Random forest classifier creates a[set](https://www.geeksforgeeks.org/set-in-cpp-stl/) of[decision trees](https://www.geeksforgeeks.org/decision-tree/) from a randomly selected subset of the training set. It is a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

*Random Forest Classifier*

Additionally, the random forest classifier can handle both classification and regression tasks, and its ability to provide feature importance scores makes it a valuable tool for understanding the significance of different variables in the dataset.

## How Random Forest Classification works

Random Forest Classification is an ensemble learning technique designed to enhance the accuracy and robustness of classification tasks. The algorithm builds a multitude of decision trees during training and outputs the[class](https://www.geeksforgeeks.org/object-oriented-programming-in-python-set-1-class-and-its-members/) that is the mode of the classification classes. Each decision tree in the random forest is constructed using a subset of the training data and a random subset of features introducing diversity among the trees, making the model more robust and less prone to overfitting.

**The random forest algorithm employs a technique called bagging (Bootstrap Aggregating) to create these diverse subsets.**

During the training phase, each[tree](https://www.geeksforgeeks.org/introduction-to-tree-data-structure-and-algorithm-tutorials/) is built by recursively partitioning the data based on the features. At each[split,](https://www.geeksforgeeks.org/how-to-split-a-string-in-cc-python-and-java/) the algorithm selects the best feature from the random subset, optimizing for information gain or Gini impurity. The process continues until a predefined stopping criterion is met, such as reaching a maximum depth or having a minimum number of samples in each leaf node.

Once the random forest is trained, it can make predictions, using each tree “votes” for a class, and the class with the most votes becomes the predicted class for the input data.

### Feature Selection in Random Forests

[Feature selection](https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) in Random Forests is inherently embedded in the construction of individual decision trees and the aggregation process.

During the training phase, each decision tree is built using a random subset of features, contributing to diversity among the trees. The process is, known as feature bagging, helps prevent the dominance of any single feature and promotes a more robust model.

The algorithm evaluates various subsets of features at each split point, selecting the best feature for node splitting based on criteria such as information gain or Gini impurity. Consequently, Random Forests naturally incorporate a form of feature selection, ensuring that the ensemble benefits from a diverse set of features to enhance generalization and reduce overfitting.