Project Design Phase-II
Data Flow Diagram

Date	20 October 2023
Team ID	Team-592697
Project Name	Diabetes Prediction Using Machine Learning
Maximum Marks	4 Marks

Data Flow:

Data Collection:

Electronic health records (EHRs): EHRs contain a wealth of data on patients' medical history, including demographics, diagnoses, procedures, medications, and laboratory results. This data can be used to train a machine learning model to identify patterns that are associated with diabetes risk.

Claims data: Claims data from insurance companies can also be used to collect data for diabetes prediction models. This data includes information on patients' medical diagnoses and procedures, as well as the cost of these services.

Wearable devices: Wearable devices such as fitness trackers and smartwatches can be used to collect data on patients' physical activity, sleep, and other health metrics. This data can be used to train a machine learning model to identify patterns that are associated with diabetes risk.

Patient surveys: Patient surveys can be used to collect data on patients' demographics, lifestyle habits, and family history of diabetes. This data can be used to train a machine learning model to identify patterns that are associated with diabetes risk.

- Data Pre-processing:

Remove outliers: Outliers are data points that are significantly different from the rest of the data. Outliers can skew the results of the machine learning model, so it is important to remove them before training the model.

Impute missing values: Missing values are data points that are not present in the dataset. Missing values can also skew the results of the machine learning model, so it is important to impute them with reasonable values before training the model.

Scale the data: Different features in a dataset can have different scales. This can make it difficult for the machine learning algorithm to learn from the data. To address this issue, the data can be scaled so that all features have a similar scale.

Transform the data: The machine learning algorithm may require the data to be in a specific format. The data can be transformed into the required format before training the model.

Model Training: In this stage, the pre-processed data is utilized to train a machine learning model, Choose a machine learning algorithm like logistic regression, neural networks.

Model Evaluation: The data is split into three sets: a train set, a validation set, and a test set. The model is trained on the train set, evaluated on the validation set, and finalized on the test set. This helps to ensure that the model is generalizable to unseen data and that the validation set is not overfitting.

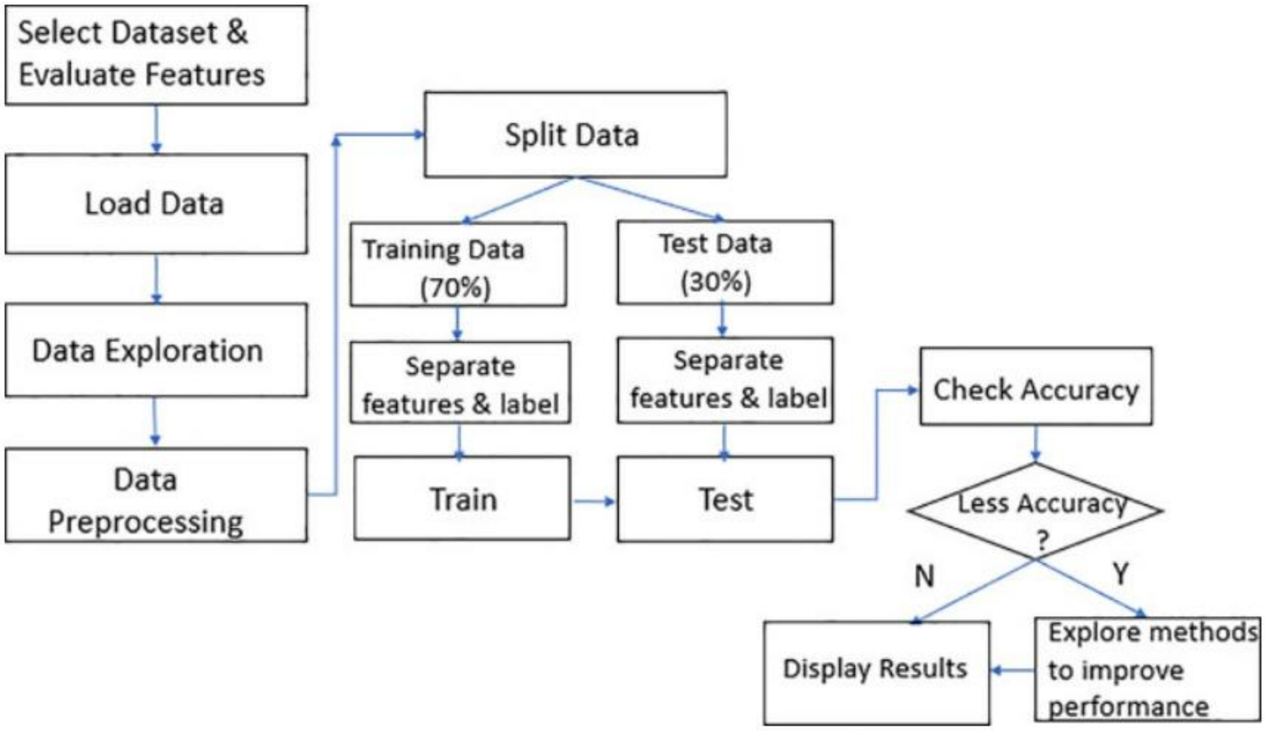
Model Deployment: Once the model is trained and evaluated, it can be deployed to production. This means making the model available to users so that they can use it to predict their diabetes risk.it can be used for web services, mobile app, embedded system

User Interaction: User interaction for a diabetes prediction model should be simple and intuitive. Users should be able to easily enter their feature data and receive a prediction of their diabetes risk. The model should also provide users with feedback on their predictions, such as explaining the factors that contributed to the prediction and providing recommendations for reducing diabetes risk.

Flow in the model:

1. The user enters their feature data, such as age, gender, weight, height, blood pressure, and blood glucose levels.
2. The model preprocesses the data by removing outliers, imputing missing values, scaling the data, and transforming it into a format that is compatible with the machine learning algorithm.
3. The model makes a prediction of the user's diabetes risk based on the preprocessed data.
4. The model provides feedback to the user on the prediction, such as explaining the factors that contributed to the prediction and providing recommendations for reducing diabetes risk.
5. The user can then decide what actions to take based on the prediction and feedback. For example, the user may decide to make lifestyle changes, such as eating a healthier diet and exercising more.
6. The user may also choose to seek preventive care from a healthcare provider.
7. The model is continuously updated with new data and improved over time to ensure that it is as accurate and reliable as possible.

Data flow diagram:


Select Dataset & Evaluate Features
Load Data
Data Exploration
Data Preprocessing
Split Data
Training Data (70%)
Test Data (30%)
Separate features & label
Separate features & label
Train
Test
Check Accuracy
Less Accuracy ?
N
Y
Display Results
Explore methods to improve performance