1. Title:

Identification of Diabetes in a person based on healthcare statistics.

2. Project Statement:

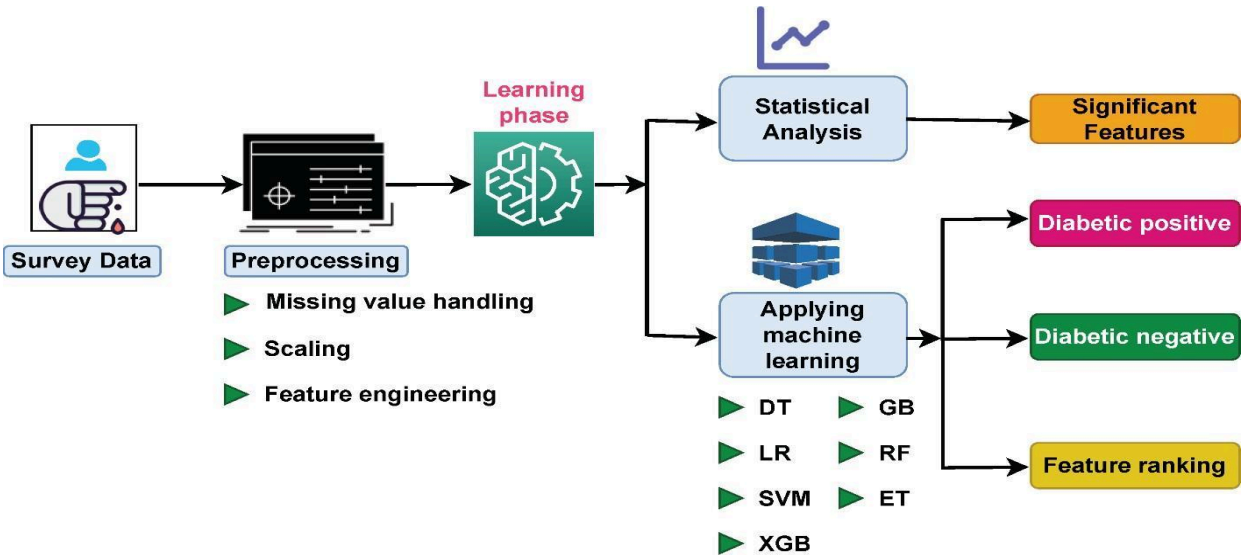
Diabetes cases over the past fifteen years have bloomed all over the world. Lifestyle plays a very important role in it. In recent years, there has been an improvement in awareness regarding the health effects of diabetes. This has led to people getting themselves tested for diabetes than they would have earlier, as its risk can be reduced if it is predicted early.

Outcomes:

The goal is to develop a model which helps better understand the relationship between lifestyle and diabetes and help predict whether a patient has diabetes, is pre-diabetic or healthy.

Modules to be implemented

1. Data Collection – Statistics of healthcare and lifestyle survey information
2. Data Exploration (EDA) and Data Preprocessing
3. Feature Selection and dimension reduction approaches
4. Build a classification model
5. Evaluation metrics
6. Presentation and Documentation

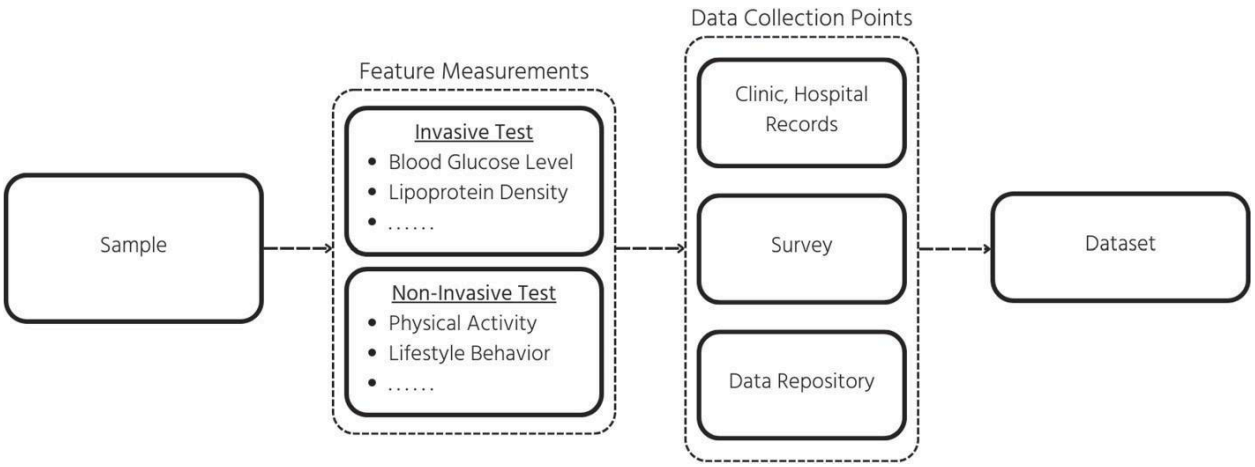


3. Week-wise module implementation and high-level requirements with output screenshots

Milestone 1: Week 1-2

Module 1: Data Collection

- Understand the objective of the use case
- Research and collect the required data



Module 2: Data Exploration and Data Preprocessing

- Data Collection and preprocessing to remove necessary discrepancies from the dataset.
- Dealing with missing data and imbalanced class issues
- Outlier observation analysis
- Encoding of the data to be performed for understanding the categorical information effectively.
- Implementing techniques such as oversampling to enable the detection models to have more samples for training.

Milestone 2: Week 3-5

Module 3: Feature Selection and dimension reduction approaches

- Selecting highly relevant features.
- To compute each feature's correlation and relative importance using specific algorithms

Module 4: Build a classification model

- Various machine learning and ensemble techniques to be employed to implement the diabetes prediction system.
- Finalize the best one based on the performance metrics of all those algorithms.

Milestone 3: Week 6-7 Module

5: Evaluation metrics

- Consider precision, recall, F1 score, AUC, and classification accuracy to evaluate various ML models.
- Comparison of the performance metrics between the model's performance at a specific threshold and models' ability to distinguish between positive negative values across all potential threshold values.
- Final selection of a model

Milestone 4: Week 8

Module 6: Presentation and Documentation

- Prepare a presentation which must include the details of the problem statement, details of the data collected, data preprocessing methods and its outcomes, model building methodology, performance metrics and recommendations based on the outcome.
- Project document which should capture the same topics mentioned above in more detailed format.

Evaluation Criteria:

Milestone 1 Evaluation (Week 1-2):

• Approval on the master dataset to be used.

• Approval on the Independent Variables to be used, based on the Univariate and Bivariate Analysis performed on the master data.

• Approval on the data preprocessing techniques.

• Approval on data treatments performed on the data.

Milestone 2 Evaluation (Week 3-5):

• Approval on different machine learning algorithms to be used on the master dataset.

Milestone 3 Evaluation (Week 6-7):

• Approval on Performance Metrics on all the built Models

Milestone 4 Evaluation (Week 8):

• Approve Final Model.

• Approve Presentation and Project Documentation.

• Approve Remediation/Action plans for the Business.

• Final Code Submissions on GitHub