The diabetes has been on the rise in recent years and significantly increased in people in their mid-life cycle. This study aims to compare the efficacy of different machine-learning models in predicting diabetes using dietary data from the National Health and Nutrition Examination Survey (NHANES) database. Machine Learning Model performance was evaluated using accuracy as the primary metrics, but other metrics will be uses are precision, F1 score, and the receiver operating characteristic (ROC) curve.

Among the current ML models Random Forest shows accuracy as 97%, KNN at 93%, CatBoost as 93% but Logistic Regression shows the accuracy as 23%. SVM and hyperparameter was taking longer compute time than usual so that will be part of next execution cycle.

By leveraging NHANES data and advanced machine-learning techniques, this study provides valuable insights into the prediction of diabetes and highlights the importance of dietary factors. The findings contribute to the growing body of research on diabetes prevention and can assist healthcare professionals in implementing targeted interventions and personalized treatment strategies for individuals at risk of diabetes.

Based on medical research, it has been established that the fasting glucose data HBA1C should be lower than 7 to avoid the risk of being diabetic whereas higher than 7 value considered as diabetic. The final dataset has imbalanced data, SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) was used to dela with it to avoid the bias. SMOTE-NC was used over SMOTE to handle categorical and numerical both the data.