# A PILOT STUDY OF DESIGN AND APPLICATION OF A MACHINE LEARNING SYSTEM FOR DIAGNOSING DIABETES


## UNIVERSITY OF ESSEX, COLCHESTER

## SCHOOL OF COMPUTER SCIENCE AND ELECTRONIC ENGINEERING

## YUNUS ALADE


**Word Count: 659 words**

## Abstract:
Diabetes is a serious blood disease that affects a large number of people around the world. It can be cause by different environmental and biological factors. Diagnosing it at its early stage is both beneficial to patients and medical practitioners as it helps to prevent or delay the onset of the disease.

This research intends to explore the appropriate machine learning model that can predict patients that are at high risk of developing diabetes, based on data provided in medical records.

## Literature Review:
Different studies have attempted to explore the role of machine learning in the early identification of patients at risk of developing diabetes. These studies have confirmed that machine learning algorithms have the potential to predict the risk of patients developing diabetes by analyzing data gotten from electronic medical records.

Decision trees is a commonly used machine learning algorithm that has been employed to predict the risk of getting diabetes. Due to its ability to process the relationship between complex data, it is regarded as one of the most effective models for prediction. Decision trees was employed in a research by Al-Rubaie et al. (2021) to forecast the likelihood that type 2 diabetes will develop in an adult population in Iraq. With an accuracy score of 0.91, the study discovered that decision trees were more accurate compared to logistic regression.

Logistic regression was used in a research by Kharroubi et al. (2020) to forecast the likelihood that type 2 diabetes will develop in a cohort of Saudi Arabian people based on the data gathered. The model had a high accuracy score of 0.87.

Random forests was used in a study by Akinwande et al. (2019) to predict the risk of type 2 diabetes in a group of people in Nigeria. The study discovered some factors that contribute to the risk of diabetes were and not limited to age, family history, obesity, high blood pressure, and physical inactivity. He labelled them as the top five factors to consider in predicting the risk of diabetes in the demography that the data was gotten from.

In general, machine learning algorithms have proven to be effective in detecting the risk of diabetes based on gathered data.

## Methodology:
The problem of identifying patients at high risk of developing diabetes is a binary classification task, we aim to classify patients into two groups: the group who are at high risk of developing diabetes (coded as “1” or “true”) and the group who are not at risk of developing diabetes (coded as “0” or “false”). Therefore, we will use a classification algorithm to solve this problem.

We need to find informative qualities that are likely to be effective predictors of the probability of acquiring diabetes in order to forecast this risk. We'll take into account a few of the following useful features:

- Age
- Gender
- Family history of diabetes
- Body Mass Index (BMI)
- Diet habits
- Blood glucose levels
- Blood pressure
- Cholesterol levels
- Smoking status
- Physical activity level
- Alcohol consumption

For this binary classification task, we will use a Decision Tree model. Decision Tree models have proven to perform well in binary classification tasks, particularly when the data is high-dimensional, and the

number of samples is eighter moderate or complex.

To test the performance of the Decision Tree model, we will employ a range of performance criteria, which are Accuracy, Precision, Recall, and F1-score, to achieve this. Cross-validation will also be used to gauge how well the Decision Tree model predicts.

## Conclusion:

In this pilot study, we propose using a Decision Tree Model to identify patients at high risk of developing diabetes because of its ability to handle large and complex data. The Decision Tree model can be trained on the provided features and evaluated using a variety of performance metrics. The 0.91 score recorded in the research by Al-Rubaie et al. (2021) gives a strong backing to the claim that Decision Tree model is suitable for predicting the risk of the diabetes in patients.

## Reference:


- Kharroubi, S., Slimani, Y., & Baghazza, M. (2020). Predicting the risk of developing type 2 diabetes using logistic regression. Journal of Medical Systems, 44(3), 1-7.
- Al-Rubaie, A. F., Al-Dubai, S. A. R., & Abood, S. H. (2021). Predicting the risk of developing type 2 diabetes using decision trees. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15(2), 557-561.
- Akinwande, O., Akinmolayan, J., & Adeyemo, T. (2019). Random forests analysis of risk factors associated with type 2 diabetes in Nigeria. Journal of Medical Systems, 43(11), 1-6.