UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SECB3203

PROGRAMMING FOR BIOINFORMATICS

PROJECT PROPOSAL

LECTURER NAME

DR. SEAH CHOON SEN

SECTION 01

GROUP 06

STUDENT NAME & MATRIC ID

RAJA NUR ALLEA DEWI MAHSURI BINTI RAJA MOHD YUSRI
A24CS0294

IRDINA SOFIA BINTI ROHAIDI
A24CS0253

NURZULAIKHA BINTI MOHD ISA
A24CS0291

TABLE OF CONTENTS

1.0 INTRODUCTION

Nowadays, chronic diseases such as heart diseases, diabetes, and cancer are the leading causes of death not only in Malaysia but also worldwide. According to Ramli, A., & Taher, S. (2008), it states that chronic diseases are the major cause of death and disability in Malaysia, with 71% of overall cases of death and 69% of the total burden of disease meanwhile it killed at least 43 million people around the world based on the data from World Health Statistics (WHO) on 2021. This study focuses on diabetes itself, which is a common disease that might occur in people nearby. Diabetes happens when the pancreas itself fails to produce enough insulin or when the blood sugar is too high and the body is not properly utilizing it, also known as Hyperglycemia. Diabetes is most likely to affect people of all ages, depending on the type. Type 1 diabetes happens when the pancreas produces insufficient insulin or no insulin at all. This type of diabetes has no cure and depends on how well the insulin. Meanwhile, type 2 diabetes happens when the body stops using insulin properly, which can lead to serious damage to the body, especially nerves or blood vessels. Gestational diabetes happens during pregnancy, where the blood sugar level is above the normal sugar level value but below the diagnostic of diabetes. The symptoms of those types can be detected at early stages by doing regular check-ups and blood tests. However, this traditional diagnostic method in most cases relies on clinical tests, which might not be accessible in certain areas, or the cost is not cheap, especially for people in urban areas. These are the reasons why it is important to use data driven methods to detect individuals at high risk of diabetes with the help of simple demographic and lifestyle information.

In this context, machine learning and data analysis are useful tools to predict health conditions and support clinical decision-making by analyzing the large volume of health-related data. By using machine learning, it can discover the patterns from the information and come up with the relationships between the factors such as body mass index (BMI), sugar level, age and smoking status. The results enable healthcare institutions to predict the risk at the earliest stage and come up with a preventive intervention. This project will apply the machine learning techniques to a dataset from Kaggle to predict the risk factors of diabetes.

1.1 PROBLEM BACKGROUND

The area of medical research has been improving due to high-end health technology as well as the advancement of data science. However, global cases due to diabetes keep rising, which causes worries if it cannot be prevented. Diabetes happens when the pancreas itself fails to produce enough insulin or when the blood sugar is too high and the body is not properly utilized it which can lead to other chronic conditions such as cardiovascular diseases, kidney failure and damage to the body's nerves. Even with that progress, diabetes remains as one of the leading causes of death and disabilities such as vision impairment.

The increasing of large datasets has made it easier to analyze the risk factors of diabetes by using machine learning. These computer-based methods can find patterns in health data that are hard to find with standard statistical methods. Machine learning models can look at a lot of different variables at once and figure out how they are related to each other in a complicated way. This study seeks to utilize supervised machine-learning algorithms on a global health dataset to forecast the probability of diabetes in individuals. The study seeks to determine the most critical factors affecting diabetes by examining various models, including Logistic Regression, Decision Tree, and Random Forest, and to assess model accuracy for early prediction. The insights derived from this analysis are anticipated to aid in the development of effective digital health applications that facilitate early intervention and enhance healthcare outcomes.

1.2 PROBLEM STATEMENT

Diabetes is a long-lasting health problem that influences millions of individuals in the world and is a severe threat when not detected and treated in the early stages. Although medical data is available, many people are not diagnosed or diagnosed at a late stage because there are no effective predictive tools. The classical approaches to the diagnosis usually base on the manual interpretation of clinical factors, which may be time-consuming and may lead to mistakes. Thus, the necessity to find more data-driven methods to outline the most important risk factors and forecast diabetes in a more efficient way is increasing.

The study will discuss the diabetes dataset to determine meaningful health indicators (including glucose levels, BMI, age, and blood pressure) that promote the risk of diabetes and create the predictive model that can help healthcare specialists make an early diagnosis and preventive decision-

making. The research will support the need to use machine learning and statistical analysis to enhance healthcare outcomes by improving the accuracy of predictions.

1.3 AIM AND OBJECTIVES

AIM

To investigate and develop a computational framework that utilizes health data to identify, analyze, and predict the chances of individuals developing diabetes to enhance preventive health care decision-making.

OBJECTIVES

- To investigate the relationships between health measures (such as glucose levels, BMI, blood pressure, and age) and the existence of diabetes through statistical and computational methods.
- To design a predictive model that combines data preprocessing, feature selection, and classification methods to identify the risk of diabetes with measurable performance results.
- To make a comparison of various machine learning algorithms to identify the most efficient and accurate machine learning model to predict diabetes.

1.4 SCOPES

DATASET

The project uses the Diabetes Health Indicators Dataset from Kaggle, derived from the Behavioral Risk Factor Surveillance System (BRFSS 2015). The dataset includes over 250,000 records representing health survey responses. It focuses on health behaviors, medical history, and demographic information that may influence the likelihood of cause diabetes. Each record includes a binary target variable “Diabetes_binary”, indicating whether an individual has diabetes (1) or not (0).

FEATURES

The Diabetes Health Indicators Dataset includes 21 features representing demographic, behavioral, and medical factors related to diabetes risk. These features cover aspects such as age, sex, BMI, blood pressure, cholesterol levels, physical activity, smoking and alcohol habits, sleep duration, and general

health status. The dataset also includes a comprehensive view of how lifestyle and health indicators can cause diabetes. This dataset has 22 columns, but this project only uses 10 columns.

TECHNIQUES

This project applies data preprocessing, exploratory data analysis, and machine learning techniques using Python to predict diabetes occurrence. Classification algorithms such as Random Forest will be trained and evaluated to determine key predictors of diabetes. The analysis and model deployment will be conducted on Microsoft Azure or AWS cloud computing platforms to ensure scalability, accessibility, and real-world cloud implementation experience.

LIMITATIONS

There are a number of limitations associated with this project. The data is based on self-reported of people, which is not always true or accurate. It further fails to provide valuable information like genetic, environmental, or medical test data, which restrains the depth of the analysis. The data was taken only once. Therefore, it cannot demonstrate the way in which the health of a person should vary as time passes. There might be a close relationship between some of the features, and this can have an impact on the learning and predictive performance of the model. Even though the advanced models of machine learning, such as the Random Forest, could produce a high accuracy, they are hard to understand and interpret. The sample also may not be quite applicable to any other population. Finally, the reliance on cloud computing providers, such as Azure or AWS, can be resource-constrained or too expensive, or inaccessible due to the internet.

1.5 CONCLUSION

In conclusion, the purpose of the project is to predict the risk of diabetes and determine the most important factors influencing its occurrence with the help of data analysis, bioinformatics, and cloud computing. The project will also show the relationship between health and lifestyle habits toward diabetes by analyzing the Diabetes Health Indicators Dataset. The implementation of Python and cloud systems like Microsoft Azure or AWS will offer an opportunity to have practical experience in data analysis and deployment. The result of this project would be helpful to create awareness of avoiding diabetes and to illustrate how data-driven practices can contribute to making people more knowledgeable in their health decisions and lifestyle choices. This project can lead to not only

understanding how health data can be used to predict diseases but also capitalizing on the discussion of how technology would aid in making people more informed in their health choices and lifestyles. Lastly, the results of the current research may help to create awareness regarding diabetes and promote more healthy lifestyles in accordance with the evidence-based findings.

REFERENCES

Ramli, A., & Taher, S. (2008). Managing chronic diseases in the malaysian primary health care - a need for change. *Malaysian family physician : the official journal of the Academy of Family Physicians of Malaysia*, *3*(1), 7–13.

World Health Organization. (2024, December 23). *Noncommunicable diseases*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases