INTRODUCTION

Type 2 diabetes is an important concept that involves a serious health condition which is said to affect millions of people around the world. This situation has the power to lead to several health complications, ranging from heart disease to kidney disease and then blindness. Timely detection and treatment of type 2 diabetes is necessary for curbing these serious hurdles.

It is a chronic metabolic disorder characterized by insulin resistance and impaired glucose regulation, resulting in elevated blood sugar levels. Unlike type 1 diabetes, which is often diagnosed in childhood and attributed to an autoimmune response, type 2 diabetes typically develops in adulthood and is associated with lifestyle factors such as obesity, physical inactivity, and genetic predisposition. This condition poses significant health risks, including cardiovascular complications, kidney disease, and nerve damage. Effective management involves lifestyle modifications, medication, and regular monitoring to control blood glucose levels and mitigate associated health risks. Understanding the intricacies of type 2 diabetes is crucial for both individuals managing the condition and healthcare professionals providing comprehensive care.

**Aims and Objectives**

The prevalence of diabetes is a major health problem worldwide. Early diagnosis of diabetes in humans is important for early intervention and effective control of the disease. Existing risk assessment tools often rely on a limited number of variables and may not have the precision required for an accurate assessment.

The goal of this research therefore is to address some root causes and challenges of diabetes type 11, optimize early detection and management and further do the analysis using other independent variables. Doing this, I will build a model, and also explore some other classification algorithms in a bit to make prediction, then conclusion and recommendation (discussion).

**Background to studies**

According to CDC report as out January, 2024, about 38 million Americans have diabetes (about 1 in 10), and approximately 90-95% of them have type 2 diabetes. Type 2 diabetes most often develops in people over age 45, but more and more [children, teens](https://www.cdc.gov/diabetes/prevent-type-2/type-2-kids.html), and young adults are also developing it.

The discuss of diabetes is broad and vast and as such it can not be exhausted. However, this project aims to develop a machine learning model for predicting the likelihood of Type 2 Diabetes based on relevant health indicators and demographic factors, providing valuable insights for early intervention and personalized healthcare. The project's effectiveness is contingent on the availability and quality of input data, potential biases in the dataset, and the inherent complexity of predicting health outcomes. Additionally, ethical considerations and the need for ongoing model validation may impact the generalizability and reliability of the predictive model.

***One way to detect the risk of developing type 2 diabetes is through predictive modelling. This predictive modelling involves three steps. The first step has to do with the data collection. The second step has to do with data preparation. And the last step is the model training and testing.***

***Our data collection involves gathering data on a large group of patients with and without type 2 diabetes.***

**RESEARCH OBJECTIVES**

The research objectives gears towards preparing a predictive model. Below is a lowdown of the various objectives:

1. Expecting the outcome of a provided dataset to determine whether an individual will encounter diabetes or otherwise.

2. Categorizing information, like discerning whether an individual possesses diabetes or not.

3. Assembling data into clusters based on likeness, such as grouping individuals according to their risk of developing diabetes.

4. Recognizing and taking out meaningful features from data, such as segregating pertinent features from dietary data related to the risk of developing diabetes.

4. we tend to understand the cause-and-effect associations between variables, such as the connection between diet and the likelihood of developing diabetes.

**Relevance**

Early detection and prediction for Type 2 diabetes risk assessment is crucial for several reasons. It enables early intervention and prevention, reducing the risk of diabetes and complications by identifying high-risk individuals at an early stage. It identifies low-risk individuals, avoiding unnecessary screening costs. Additionally, predictive modeling helps recognize those at risk of complications, enhancing interventions and reducing associated risks. It streamlines healthcare efficiency and cost management, conducting risk assessments swiftly and efficiently. Integration with healthcare systems and data sources offers a holistic view of health, promoting better coordination and information sharing among providers. Lastly, it contributes to health equity by addressing disparities related to socioeconomic, racial, and ethnic factors, ensuring widespread access to information and interventions.

**LITERATURE REVIEW**

Type 2 diabetes mellitus (T2DM) is a prevalent and becoming more chronic metabolic disorder characterized by insulin resistance and impaired glucose regulation. Data has shown that it is escalating globally that in every one in ten individuals has it. Therefore, there is a growing emphasis on preventive measures and early identification of individuals at risk. Predictive modeling, utilizing various statistical and machine learning approaches, has emerged as a promising tool for assessing the risk of developing diabetics type 2. This literature review aims to explore and synthesize existing research on the application of predictive models in identifying individuals at risk of T2DM

A comprehensive search was conducted across various academic databases, using keywords such as "predictive model," "risk assessment," and "Type 2 diabetes." Articles published in the Journal of Diabetology and metabolic syndrome (2021). The title of the article is **“Machine learning and deep learning predictive models for type 2 diabetes: a systematic review".** It found out that predictive models can be effective in predicting the risk of developing type 2 diabetes, but there are some limitations to the existing models. One key finding is that the accuracy of predictive models can vary depending on the data that is used to train the model. Another finding is that predictive models can be improved by incorporating new data sources, such as wearable devices and social media data.

It explored machine learning techniques in diabetes, yet with a substantially different focus. One of the researchers, Sambyal et al. conducted a review on microvascular complications in diabetes. This review included 31 studies classified into three groups according to the methods used: statistical techniques, machine learning, and deep learning. The researchers concluded that machine learning and deep learning models are more suited for big data scenarios. Also, they observed that the combination of models produced improved performance.

Another researchers Islam et al. conducted a review with meta-analysis on deep learning models to detect diabetic retinopathy (DR) in retinal fundus images. This review included 23 studies, out of which 20 were also included for meta-analysis. For each study, the researcher identified the model, the dataset, and the performance metrics and concluded that automated tools could perform DR screening.

Chaki et al. reviewed machine learning models in diabetes detection. The review included 107 studies and classified them according to the model or classifier, the dataset, the features selection with four possible kinds of features, and their performance. The researchers found that text, shape, and texture features produced better outcomes. Also, they found that DNNs and SVMs delivered better classification outcomes, followed by RFs.

From the above, it shows that researchers have employed diverse predictive models, including machine learning algorithms, statistical regression, and data mining techniques. Machine learning, such as support vector machines, decision trees, and neural networks, has gained prominence for its ability to discern complex patterns within large datasets. Selecting an algorithm or Machine learning algorithm to fit into a data follow an existing structure. The classifications of machine learning algorithms are described below. There are many machine learning techniques that fall into three categories: supervised learning, unsupervised learning, and semi-supervised learning. However, I will only be discussing few of them randomly.

1. Simple linear regression is a statistical method used to model the relationship between a single independent variable and a dependent variable. It assumes a linear relationship between the two variables, meaning that changes in the independent variable are associated with a constant change in the dependent variable.

In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero. the formula is

y = β0 +β1x+ε.

where y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).

B0 is the intercept, the predicted value of y when the x is 0.

B1 is the regression coefficient – how much we expect y to change as x increases.

x is the independent variable (the variable we expect is influencing y).

e is the error of the estimate, or how much variation there is in our regression coefficient estimate

2. Logistic regression: This is a type of regression analysis that is used to predict the probability of a binary outcome, such as whether or not a patient develops a disease. Logistic regression models can be simple or complex, and they can be built using a variety of methods. One popular method is the maximum possibility method which uses statistical techniques to find the values of the model's parameters that maximize the likelihood of the observed data.

Logistic regression models are created and used by first collecting data and selecting the variables. Next, the data is split into training and test sets, and the model is built on the training set. Finally, the model is validated on the test set, and the accuracy of the model is assessed.

3. Decision Trees: This is a kind of model that is utilized to illustrate decisions and their feasible results in a tree-like hierarchy. It is often done in a bid to predict the consequence of a conclusion relating to the characteristics of the given problem. Furthermore, decision trees are well-known as a result of the fact that they are not difficult to understand and it is not necessary to use so much data before they can be built.

4. Support Vector Machines (Svms): SVMs are a type of machine learning algorithm that is particularly well-suited for classification problems. SVMs find a hyperplane that separates the data into two classes and then maximizes the distance between the two classes. One of the main advantages of SVMs is that they are very effective at finding complex decision boundaries. On the other hand, they can be slow to train, and they are sensitive to the choice of parameters.

5. Neural Networks: Neural networks are stimulated by the pattern of the human brain, and they are used to solve complex problems like image recognition and natural language processing. One of the main benefits of neural networks is that they can learn to make revelations on their own, and this can be done without any precise programming. Nonetheless, one problem with this is that it can be difficult to interpret and elucidate.

6. Random Forests: A random forest is a type of ensemble method that combines multiple decision trees to improve the accuracy of predictions. The key advantage of random forests is that they can handle large amounts of data and are more robust to outliers. The main drawback is that they can be slow to train and require a lot of memory.

The application of predictive analysis in Type 2 diabetes offers substantial benefits, including early detection, personalized treatment plans, and resource optimization in healthcare settings. Future research should focus on refining existing models, validating predictive accuracy across diverse populations, and establishing clear clinical guidelines for implementing predictive analytics in routine healthcare practices.

**CHAPTER TWO**

**OVERVIEW OF TYPE 2 DIABETES**

Type 2 diabetes is a chronic metabolic condition that affects the way the body uses glucose, a type of sugar found in the blood. People with type 2 diabetes have trouble using insulin, a hormone that helps regulate blood sugar levels. This causes the blood sugar to rise, which can lead to serious health problems if left untreated.

**THE RISK FACTORS FOR TYPE 2 DIABETES**

Lifestyle factors that can increase the risk of type 2 diabetes include being overweight or obese, not getting enough physical activity, and eating a diet high in fat and sugar. There are also genetic factors that can increase the risk of developing type 2 diabetes. Some of these genetic factors are related to race and ethnicity, and others are related to specific genes that can affect the way the body uses insulin.

**SYMPTOMS OF TYPE 2 DIABETES**

The main symptoms of type 2 diabetes include increased thirst and hunger, frequent urination, blurred vision, unexplained weight loss and tiredness. Some people with type 2 diabetes may also experience skin infections, sores that do not heal, and numbness or tingling in the hands and feet. These symptoms are caused by the high blood sugar levels that occur in people with type 2 diabetes.

**HOW TYPE 2 DIABETES IS DIAGNOSED AND TREATED**

To diagnose type 2 diabetes, it is expected of a healthcare provider to conduct a test to know the high blood sugar levels of the patient. If the levels are much, the healthcare provider may surmise that the victim is battling with the disease called type 2 diabetes. The verification of the diagnosis will then require the healthcare provider to protect the patient by suggesting changes in diet and activity levels, as well as drugs to be admitted to control blood sugar levels. In some cases, the patient may also need to take insulin injections. The purpose of the treatment plan is to keep blood sugar levels at bay and prevent complications.

In summary, research has shown that close to thirty million people in the United States are victims of this disease. Worse still, the number of victims is rising rapidly every day. In addendum, type 2 diabetes is connected to other diseases, such as kidney disease, vision loss, heart disease and stroke. People are advised to get regular checkups to manage their conditions and keep the risks of complications at bay.

**PREVALENCE OF TYPE 2 DIABETES**

The prevalence of Type 2 diabetes varies in terms of the population and country. However, in general, it is a very widespread ailment. In the United States, for example, the prevalence of Type 2 diabetes is estimated to be about 10% of the adult population. The prevalence is even higher in specific ethnical and racial groups, such as African Americans and Hispanic/Latino Americans. Type 2 diabetes is also more prevalent in older people, obese or have other risk factors such as a family history of diabetes.

In Europe, the prevalence of diabetes varies widely from 4% in Sweden to 14% in Malta. In Asia, the prevalence ranges from 4% in Japan to 11% in Thailand. In the Middle East and North Africa, the prevalence is estimated to be around 6%. And in Latin America, the prevalence ranges from 4% in Chile to 12% in Mexico. This data only unravels a general overview, whereas the actual prevalence of diabetes may differ within each country and region. For example, in Canada, the prevalence of diabetes is estimated to be around 6%, while in South Africa, it is estimated to be around 7%. In Australia, the prevalence is estimated to be around 8%. It should be noted that these numbers are estimates and may not be completely accurate.

Globally, the prevalence of diabetes is most common among 65-year-old-and-above people, and it is less common among people between 18-44 years. Under each age group, there are also discrepancies in the preponderance of diabetes between men and women. For example, the bulk of diabetes is higher in men than in women aged 45-64.

The prevalence of diabetes also differs by socioeconomic status. In general, people with lower socioeconomic status have a higher prevalence of diabetes than those with higher socioeconomic status. This may be due to factors such as diet, access to healthcare, and other social determinants of health. If you're interested, I can give you more information on how socioeconomic status affects the risk of diabetes.

In the United States, studies have shown that people living in poverty are about twice as likely to have diabetes as those with higher incomes. In Canada, the prevalence of diabetes is about 30% higher in low-income groups compared to high-income groups. In Australia, people in the lowest socioeconomic group have about a 50% higher prevalence of diabetes than those in the highest socioeconomic group. There are also similar patterns in other countries, although the specific numbers may vary.

**IMPORTANCE OF EARLY RISK ASSESSMENT**

In essence, early risk assessment can go a long way to protect lives. At the identification of individuals who are at high risk of having diabetes, healthcare professionals can step in early to curb the disease from developing in the first place.

Their involvement or intervention can impede numerous severe inconveniences, including heart disease, stroke, kidney failure, vision loss and limb amputation. Even if diabetes is not controlled early recognition and treatment are significant because they help to restrict the flow of the disease and then enhance the value of life for the victims of type 2 diabetes. Additionally, the application of this strategy will generate considerable cost savings for the healthcare system in the long run.

When the flow of diabetes is hindered, the costs relating to the treatment and management of the disease can be lowered. As a result, resources that would have been used to treat the prevented diseases will be applied to treating other conditions of people. In a nutshell, early risk assessment helps to save costs, and it is the surest way of protecting people from the condition of type 2 diabetes.

There is also a credible disagreement to examine here. Some scholars might insist that it is heinous to abstain from information about a person's risk of getting diabetes, as this could lead to them creating lifestyle options that put them in a bigger danger of generating the disease. By giving people information regarding their risk, they can make knowledgeable conclusions about their health and take the essential measures to lessen their risk. In some cases, this might mean making changes to their diet, exercise habits or other lifestyle factors. Eventually, early risk examination provides people with the ability to make decisions about their health.

One final point to consider is the potential impact of predictive modelling on public health. If we can identify people at high risk of developing diabetes, we can develop targeted interventions to help reduce the incidence of the disease in the population as a whole. This could lead to a significant decrease in the number of people who develop diabetes, which would be a major public health victory. By improving the health of individuals, we can improve the health of society as a whole.

Additionally, predictive modelling can be a powerful tool for research. By recognizing people at high risk of developing diabetes, we can study their genetics, lifestyle, environment and other factors to better understand what causes the disease. This can lead to the development of new treatments and prevention strategies. So, predictive modelling not only has the potential to improve individual health, but it could also lead to a better understanding of diabetes as a whole.

**Relevant Predictive Modelling Techniques.**

There are so many techniques that can be used for predictive modelling, but our focus will be on a few of the most relevant ones. One common technique is logistic regression, which is a statistical method for analyzing data and making predictions. Another popular technique is decision trees, which are used to make predictions based on a series of "if-then" rules. Other common techniques include support vector machines, neural networks and random forests.

1. Logistic regression: This is a type of regression analysis that is used to predict the probability of a binary outcome, such as whether or not a patient develops a disease. Logistic regression models can be simple or complex, and they can be built using a variety of methods. One popular method is the maximum possibility method which uses statistical techniques to find the values of the model's parameters that maximize the likelihood of the observed data.

Logistic regression models are created and used by first collecting data and selecting the variables. Next, the data is split into training and test sets, and the model is built on the training set. Finally, the model is validated on the test set, and the accuracy of the model is assessed.

2. Decision Trees: This is a kind of model that is utilized to illustrate decisions and their feasible results in a tree-like hierarchy. It is often done in a bid to predict the consequence of a conclusion relating to the characteristics of the given problem. Furthermore, decision trees are well-known as a result of the fact that they are not difficult to understand and it is not necessary to use so much data before they can be built.

3. Support Vector Machines (Svms): SVMs are a type of machine learning algorithm that is particularly well-suited for classification problems. SVMs find a hyperplane that separates the data into two classes and then maximizes the distance between the two classes. One of the main advantages of SVMs is that they are very effective at finding complex decision boundaries. On the other hand, they can be slow to train, and they are sensitive to the choice of parameters.

3. Neural Networks: Neural networks are stimulated by the pattern of the human brain, and they are used to solve complex problems like image recognition and natural language processing. One of the main benefits of neural networks is that they can learn to make revelations on their own, and this can be done without any precise programming. Nonetheless, one problem with this is that it can be difficult to interpret and elucidate.

4. Random Forests: A random forest is a type of ensemble method that combines multiple decision trees to improve the accuracy of predictions. The key advantage of random forests is that they can handle large amounts of data and are more robust to outliers. The main drawback is that they can be slow to train and require a lot of memory.

**CHAPTER THREE: METHOD AND DATA**

In this chapter, I shall be discussing the process developments and strategies for detecting diabetes type 2 and explanation of the dataset according to their variables as predictors and target.

**A. Data Collection and Preprocessing**

The first step in predictive modelling is data collection. For this project, my source dataset is from https://www.kaggle.com/code/mvanshika/diabetes-prediction. It involves gathering data from a variety of sources, such as medical records, insurance claims and patient surveys. After the data is collected, it needs to be preprocessed. The process involves cleaning and transforming the data to make it ready for analysis. The most common preprocessing steps are data cleaning, data integration and data reduction.

About the Dataset

The dataset is a composition of different categorical variables. Each variable has its own role on the diabetes effect in the human body. The variables are as follows;

* Pregnancies: Number of times a woman has been pregnant
* Glucose: Plasma Glucose concentration of 2 hours in an oral glucose tolerance test
* Blood Pressure: Diastolic Blood Pressure (mm hg)
* Skin Thickness: Triceps skin fold thickness(mm)
* Insulin: 2-hour serum insulin (mu U/ml)
* BMI: Body Mass Index ((weight in kg/height in m) ^2)
* Age: Age(years)
* DiabetesPedigreeFunction: scores likelihood of diabetes based on family history)
* Outcome: 0(doesn't have diabetes) or 1 (has diabetes)

1**. Data Cleaning:** Data cleaning is the process of identifying and correcting errors and inconsistencies in the data. This can involve removing duplicates, fixing missing values and normalizing data. It is an iterative process involving techniques that take care the data. The goal is to make sure the data is accurate and complete before it is used for analysis.

2. **Data integration**: Data integration is the process of combining data from different sources into a single, unified dataset. This can be challenging because the data may be stored in different formats or use different coding systems. However, it is an important step to have a complete dataset for analysis.

3. **Data Reduction:** This involves reducing the size of the dataset by removing irrelevant or redundant data.

**4 Data validation**

Data validation is the process of ensuring that the dataset that is to be used for analysis is accurate, consistent, and meets certain predefined criteria. Therefore, the goal of data validation is to maintain data integrity and reliability, preventing errors, inaccuracies, and inconsistencies that can impact the quality of information within a system.

Data validating is essential as it helps to improve building models. Data validation entails three different processes, namely; Checking if the data is incorrect, making sure the data structure has not been modified, ensuring that predictions from our new datasets continue to match those from our previous training datasets.

**Testing and training of data**

Testing and training of dataset is very essential in build model. It is most probably one of the fundamental steps to take. It is the process of splitting data into training and testing sets is essential for assessing the performance and generalization ability of machine learning models. This practice involves dividing the dataset into two subsets: one for training the model and another for evaluating its performance on unseen data. It is important to note that most time data can perform poorly in machine learning. This could be due to noisy or corrupted data such as outliers, inconsistencies in the data can also negatively impact performance. Others could be data imbalance, high dimensionality and data drift. These can cause underfitting or overfitting in the training of data. When a model is to simple to and can not capture the underlying patterns in the data, thereby resulting to poor performance on both training and testing is called underfitting. On the other hand, when a model is too complex and learn the noise in the training data, thereby performing well on the training data but poorly on the new and unseen data is referred to as Overfitting.

**Data Visualization**

Data visualization is the representation of data in graphical or visual formats such as charts, graphs, and maps, to facilitate understanding, analysis, and interpretation of patterns, trends, and insights within the data. Visualizing data help us understand the data. It is a crucial component of data analysis and communication, allowing individuals to comprehend complex information more easily than they would through raw data. Most statistical tool for data visualization is the Power Bi or Tableau.

**Model Performance**

There are a few main metrics that are used to measure the performance of a predictive model. The most common are accuracy, precision, recall and F1 score.

For better understanding, accuracy is simply the proportion of correct predictions out of the total number of predictions. While it's a useful metric, it can be misleading if the dataset is unbalanced (i.e., if some classes are much more common than others). Precision is the proportion of positive predictions that are actually correct. It's often used in conjunction with recall, which is the proportion of positive examples that are correctly identified by the model. The F1 score combines both precision and recall into a single metric that balances their relative importance. It is a useful metric because it can be used to compare models with different levels of precision and recall.

Another popular metric is the area under the receiver operating characteristic (ROC) curve, or AUC. The ROC curve plots the true positive rate (i.e., recall) against the false positive rate for different thresholds of the model. The AUC is the area under this curve and ranges from 0 to 1. A model with a higher AUC is considered better.

Another vital metric is the confusion matrix. A confusion matrix is a table that shows the number of true positives, false positives, true negatives, and false negatives for a given dataset. This is useful for visualizing the performance of a model and identifying potential sources of error.

Another set of metrics is the kappa statistics and Cohen's kappa. These metrics measure the agreement between a model's predictions and the "ground truth" labels in the data. The kappa statistics range from -1 to 1, with higher values indicating better agreement. Cohen's kappa adjusts for agreement due to chance, and is also calculated between -1 and 1.

Wrapping up, there are also a few more metrics that are often used in specific applications. In regression problems, the root mean squared error (RMSE) is often used to measure the error of the model. In text classification problems, metrics like recall at k and the mean reciprocal rank are often used. These metrics measure the performance of the model when it is required to return a fixed number of results (k) or the best results first.

**Model Evaluation**

Once a model has been built, it is important to evaluate its performance to see how well it is working. Many different metrics can be used for model evaluation, such as accuracy, precision and recall.

1. **Accuracy**: Accuracy is a measure of how well the model's predictions match the actual values in the dataset. It is calculated by dividing the number of correct predictions by the total number of predictions. The higher the accuracy, the better the model is at predicting the outcome.

2. **Precision:** Precision is a measure of how many of the model's positive predictions are incorrect.

3. **Recall:** This is also known as sensitivity. It is a measure of how many of the actual positive cases were correctly identified by the model. It is calculated by dividing the number of correct positive predictions by the total number of actual positive cases. To sum up, accuracy, precision and recall are all important metrics for evaluating the performance of a predictive model.

**CHAPTER FOUR**

This chapter will perform the analysis of the dataset. The analysis will be performed on the publicly available Prima Indians Diabetes dataset. It contains data on diabetes among 768 women aged 21-81. The dataset was originally published by the National Institute of Diabetes and Digestive and Kidney Diseases.

 In the sample group, there was a very high ratio of diabetes, as many as 268 cases, which is almost 35% of the entire set. This is a very high result, compared to only about 10% of adult women diagnosed with diabetes years back.

The dataset contains the binary variable ‘Outcome’ (1-diabetes, 0-no diabetes) and 8 health-related variables:

* Past pregnancies
* Blood pressure
* Glucose level
* The thickness of the skin fold of the triceps
* Insulin level
* BMI Index (Body Mass Index)
* Family history of diabetes
* Age

The dataset has 768 observations and 8 variables.















