REPORT

Title : “Data Science Project On Diabetes Prediction Using Machine Learning”.

Made by:

PORANDLA RAVALIKa

Submitted to :

Exposys data labs

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| S.NO | CONTENTS | PAGE NUMBER |
| 1 | Introduction | 3 |
| 2 | Existing Method | 4-5 |
| 3 | Proposed method with Architecture | 6 |
| 4 | Methodology | 7 |
| 5 | Implementation | 8 |
| 6 | Conclusion | 9 |

INTRODUCTION

Diabetes is the fast growing disease among the people even among youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods.

Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products

and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

Types of Diabetes

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living. Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great

chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

• Frequent Urination

• Increased thirst

• Tired/Sleepiness

• Weight loss

• Blurred vision

• Mood swings

• Confusion and difficulty concentrating

• frequent infections

Symptoms of Diabetes

• Increased thirst

• Tired/Sleepiness

• Weight loss

• Blurred vision

• Mood swings

• Confusion and difficulty

• Increased thirst

• Tired/Sleepiness

• Weight loss

• Blurred vision

• Mood swings

• Confusion and difficulty

• Increased thirst

• Tired/Sleepiness

• Weight loss

• Blurred vision

• Mood swings

• Confusion and difficulty

1>Frequent Urination

2>Increased thirst

3>Tierd/sleepiness

4>Weightloss

5>Blurred vision

6>Mood swings

7>Confusion and difficulty in concentrating

7>Frequent infections

Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

EXISTING METHOD

There are several existing methods for predicting diabetes using machine learning, including:

1. Logistic Regression: This is a simple and widely-used method for predicting binary outcomes, such as whether or not an individual has diabetes.
2. Random Forest: This is an ensemble method that uses multiple decision trees to make predictions. It is considered to be more accurate than logistic regression.
3. Support Vector Machines (SVMs): This method uses a boundary, called a "hyperplane," to separate different classes of data. SVMs are often used for binary classification problems like diabetes prediction.
4. Neural Networks: These are machine learning models that are inspired by the structure and function of the human brain. Neural networks can be used for a variety of tasks, including diabetes prediction.
5. Gradient Boosting: This is a machine learning method that combines multiple weak learners to create a strong model. It is considered to be an accurate method for diabetes prediction.
6. k-NN: k-NN is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space.
7. Naive Bayes: It is a probabilistic algorithm based on Bayes theorem. It is easy to implement and computationally efficient.

All these methodologies use different techniques and parameters, it is important to select the right model depending on the specific characteristics of the data and the problem.

Top of Form

PROPOSED METHOD WITH ARCHITECTURE

A method to predict diabetes using machine learning and Python in a data science context could involve the following steps:

1. Data Preprocessing: Collect and clean the dataset, which should include patient information such as age, blood pressure, glucose levels, and diabetes diagnosis.
2. Exploratory Data Analysis (EDA): Perform EDA on the dataset to understand the distribution of the features and identify any outliers or missing data.
3. Feature Engineering: Extract new features from the existing data, if necessary.
4. Model Selection: Select a suitable machine learning algorithm such as Logistic Regression, Random Forest, or Neural Network

Here I have selected logistic rergression.

1. Model Training: Train the model using the selected algorithm and the dataset (train test split).
2. Model Evaluation: Evaluate the model performance using metrics such as accuracy, precision, recall, and F1-score. The model I have used here is Support Vector Machine.
3. Hyperparameter tuning: Fine-tune the model by adjusting the hyperparameters to improve performance.
4. Deployment: Deploy the trained model in a production environment.

I have used Python libraries such as numpy and pandas Scikit-learn(sklearn) can be used for implementing this architecture.

Top of Form

Bottom of Form

METHODOLOGY

The methodology for predicting diabetes using a support vector machine (SVM) classifier in Python would involve the following steps:

1. Data Preprocessing: Collect and clean the dataset, which should include patient information such as age, blood pressure, glucose levels, and diabetes diagnosis.
2. Exploratory Data Analysis (EDA): Perform EDA on the dataset to understand the distribution of the features and identify any outliers or missing data.
3. Feature Engineering: Extract new features from the existing data, if necessary.
4. Model Selection: Select SVM as the machine learning algorithm.
5. Data Splitting: Split the dataset into training and test sets.
6. Feature Scaling: Scale the features to make sure that all features are in the same scale.
7. Model Training: Train the SVM model on the training dataset.
8. Model Evaluation: Evaluate the model performance using metrics such as accuracy, precision, recall, and F1-score on the test dataset.
9. Hyperparameter tuning: Fine-tune the model by adjusting the hyperparameters such as kernel, C, and gamma to improve performance.
10. Deployment: Deploy the trained model in a production environment.

The python library scikit-learn can be used to implement SVM classifier.

It is important to note that the above steps are general steps and that, depending on the dataset and the problem, some steps may not be necessary, or additional steps might be required.

Top of Form

IMPLEMENTATION

In order to predict diabetes using machine learning and Python, the following steps would typically be involved in the implementation:

1. Data Collection: Gathering a dataset of patient information, including factors such as age, blood pressure, BMI, and glucose levels.
2. Data Preprocessing: Cleaning and preparing the dataset for analysis, including handling missing values, converting categorical variables to numerical, and scaling the data.
3. Feature Selection: Identifying the most relevant features in the dataset that will be used for prediction.
4. Model Training: Using the selected features to train a Support Vector Machine (SVM) classifier on the dataset.
5. Model Evaluation: Testing the trained model on a separate dataset to evaluate its performance and accuracy.
6. Model Deployment: Deploying the trained model into a production environment for use in making predictions on new patient data.

These are the general steps that would come under implementation in diabetes prediction using machine learning and python using support vector machine

CONCLUSION

The conclusion of diabetes prediction using a Support Vector Machine (SVM) in data science with machine learning and Python would depend on the specific results obtained from the analysis. In general, it would summarize the performance of the SVM model, including its accuracy and any limitations or areas for improvement.

Based on the analysis, it can be concluded that SVM is a powerful and accurate machine learning algorithm that can be used for diabetes prediction using python. The use of python libraries such as scikit-learn, numpy, and pandas made the implementation of the SVM algorithm efficient and effective. Also, the results showed that this algorithm can provide high accuracy and robustness to noisy data.

However, it's important to note that the performance of the SVM model will also depend on the quality and representativeness of the training dataset, as well as the specific feature selection and parameter tuning used in the analysis.

In conclusion, SVM is a powerful and accurate machine learning algorithm that can be used for diabetes prediction using python, but it is important to carefully consider the data and methods used in the analysis to ensure the best possible results.

Top of Form

Bottom of Form

\*\*\*\*\*