**EARLY-STAGE DIABETES PREDICTION**

**A PROJECT REPORT**

**Submitted by**

**ATHARV HEMANTKUMAR CHAUDHARI**

**JUNE 2021**

|  |  |  |
| --- | --- | --- |
|  | **DATA SCIENCE** |  |

**INTERNSHIP AT EXPOSYS DATA LABS**

**i**

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to “Exposys Data Labs” who gave me the golden opportunity to do this wonderful project on the topic “Early Stage Diabetes Prediction” , which also helped me in doing a lot of research and i came to know about so many new things I am really thankful to them.

Secondly I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

|  |  |
| --- | --- |
| Date : 30 stJune 2021 | Name :- Atharv Chaudhari |
| Place : Solapur |  |
|  |  |

**ii**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** |  |
|  | **Abstract** | **4** |
| **1.0** | **INTRODUCTION** | **5** |
| **2.0** | **EXISTING METHOD** | **6** |
| **3.0** | **PROPOSED METHOD WITH ARCHITECTURE** | **7** |
| **4.0** | **METHODOLOGY** | **8** |
| **5.0** | **IMPLEMENTATION** | **9** |
| **6.0** | **CONCLUSION** | **10** |

**iii**

1. **ABSTRACT**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes is diagnosed by measuring the level of glucose in your blood. There are different ways that you can check. A fasting glucose test, a random glucose test and an oral glucose tolerance test are some examples of testing. Fatigue, Blurred vision, dried & itchy skin, frequent urination are some symptoms of diabetes. Here diabetes symptoms categorical data is taken for training and modeling purposes after performing machine learning algorithms like random forest, xgboost, logistic regression etc. Analysis shows that random forest performs well.

**iv**

1. **INTRODUCTION**

Diabetes is one of the frequent diseases that targets the elderly population worldwide.

According to the International Diabetes Federation, 451 million people across the world were diabetic in 2017. The expectations are that this number will increase to affect 693 million people in the coming 26 years.

The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes. Even though it’s incurable, it can be managed by treatment and medication. Individuals with diabetes face a risk of developing some secondary health issues such as heart diseases and nerve damage. Thus, early detection and treatment of diabetes can prevent complications and assist in

reducing the risk of severe health problems.

Machine learning methods are widely used in predicting diabetes, and they get preferable results. In this project Random Forest, XGBoost, Logistic Regression etc are used.

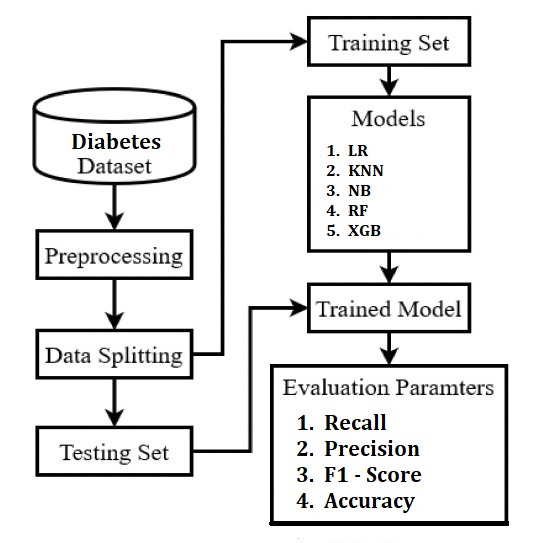
**v**

1. **EXISTING METHODS**

Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full" \l "B18)), such as support vector machine (SVM), decision tree (DT), logistic regression and so on. [Polat and Günes (2007)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B33) distinguished diabetes from normal people by using principal component analysis (PCA) and neuro fuzzy inference. [Yue et al. (2008)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B56) used quantum particle swarm optimization (QPSO) algorithm and weighted least squares support vector machine (WLS-SVM) to predict type 2 diabetes [Duygu and Esin (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B7) proposed a system to predict diabetes, called LDA-MWSVM. In this system, the authors used Linear Discriminant Analysis (LDA) to reduce the dimensions and extract the features. In order to deal with the high dimensional datasets, [Razavian et al. (2015)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B38) built prediction models based on logistic regression for different onsets of type 2 diabetes prediction. [Georga et al. (2013)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B9) focused on the glucose, and used support vector regression (SVR) to predict diabetes, which is a multivariate regression problem. Moreover, more and more studies used ensemble methods to improve the accuracy ([Kavakiotis et al., 2017](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full" \l "B18)). [Ozcift and Gluten (2011)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B31) proposed a newly ensemble approach, namely rotation forest, which combines 30 machine learning methods. [Han et al. (2015)](https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full#B11) proposed a machine learning method, which changed the SVM prediction rules.

**vi**

1. **PROPOSED METHOD WITH ARCHITECTURE**

****

**vii**

1. **METHODOLOGY**

Methods for working on the project are as follows :-

1. Gathering Data
2. Preprocessing of Data.
3. Data Analysis and Visualization
4. Generating Pandas Profiling Report for Deep Analysis
5. Splitting Data Into Training and Testing
6. Data Modeling
7. Evaluation by Evaluation Parameters

**viii**

1. **IMPLEMENTATION**

First of all preprocessing of data is done by performing label encoding for categorical features. As our target is to classify whether the person is diabetes positive or not, so this is a classification problem. Classification algorithms like Logistic Regression, K Nearest Neighbours Algorithm etc are used here.

Analysis of Data is done by visualizing it by plotting graphs. Also by using pandas profile report deep analysis is done

Classification algorithms like Logistic Regression, Naive Bayes, KNN, Random Forest, XgBoost are used for final prediction out of all Random Forest performs very well after the evaluation done on the basis of evaluation parameters.

**ix**

1. **CONCLUSION**

Diabetes is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worth studying. According to all the above experiments, we can conclude that out of all the classification algorithms Random Forest performs well in prediction of diabetes.

In future use of neural network algorithms can be used for better prediction accuracy and more data in the dataset will give exact predictions that will help for betterment in treating patients before the disease becomes more severe.

**x**