Predictive Modeling of Diabetes Risk Using Machine Learning on Health Marker Datasets: An Analytical Approach.

Dondapati.Sriramreddy [1, a)], Kolusu.Mahesh Babu [2, b)], Dandu Vamsi[3, c)], Thamarala. Sudheer Kumar[4, d)],Yadavalli. Sai Siddhardh[5, e)] ,and Aman Deep [6, f)]

[1,2,3,4,5,6]*Department of Computer Science and Engineering,*
Lovely Professional University,
Jalandhar, 140011,
Punjab, India

Corresponding author: [f)]*aman.29605@lpu.co.in*
Contributing authors: [b)]*maheshbabut1095@gmail.com* ,
[c)]*vamsidandu789@gmail.com* , [a)]*sriramreddy0902@gmail.com,* [e)] *saisiddhardha789@gmail.com,*
[d)],*sudheerthamarla123@gmail.com*

Abstract

Diabetes is a chronic disease that affects how the body processes glucose, which is the primary source of energy for the cells. There are two main types of diabetes: type 1 diabetes, which is caused by the immune system attacking the cells in the pancreas that produce insulin, and type 2 diabetes, which is caused by the body becoming resistant to insulin or the pancreas not producing enough insulin. In this study, we have selected the Health Indicator Diabetes Dataset from Kaggle. The dataset allows us to illustrate which indicator which is relevant to cause diabetes in the common people with different economic, education, and other kinds of day-to-day habit which may cause or impact the body to trigger to our immune system to stop release of the insulin which cause us to become Diabetes.Our focus on this project to build the Machine Learning based system which can detect the early stage diabetes. For this purpose, from the dataset we have consider the three samples, which was given to us. We have selected only one sample which has around 70,692 patients, as it is balanced dataset. After performing cleaning, analyzing the dataset and we have built different model and evaluate their performance in terms of accuracy and time taken for building. We have evaluated and discuss each model with their pros and cons. In conclusion, we have discussed the models which we built for early detection of diabetes in everyone life. Outcomes of our model to show a person is becoming diabetes if the given indicator is no change in our day-to-day lifestyle.

INTRODUCTION

Diabetes is a chronic disease that affects how the body processes glucose, which is the primary source of energy for the cells. There are two main types of diabetes: type 1 diabetes, which is caused by the immune system attacking the cells in the pancreas that produce insulin, and type 2 diabetes, which is caused by the body becoming resistant to insulin or the pancreas not producing enough insulin. Diabetes has become increasingly common in recent years, and it is estimated that millions of people worldwide are affected by this disease. In the United States, according to the Centers for Disease Control and Prevention (CDC), as of 2018, 34.2 million Americans have diabetes, and 88 million have prediabetes. However, the CDC estimates that many people with diabetes or prediabetes are unaware of their risk. Specifically, they estimate that 1 in 5 people with diabetes and roughly 8 in 10 people with prediabetes are unaware of their condition. There are several reasons why people may be unaware of their risk for diabetes or prediabetes. One reason is that the early symptoms of diabetes can be mild or even absent, which can make it difficult to detect the disease in its early stages. Some of the early symptoms of diabetes include increased thirst, frequent urination, fatigue, blurred vision, and slow wound healing. Another reason why people may be unaware of their risk for diabetes or prediabetes is that they may not undergo regular health checkups or screenings. Regular health checkups and screenings are important for detecting diabetes and other health conditions in their early stages when they are more treatable. In conclusion, diabetes is a chronic disease that affects millions of people worldwide, and its prevalence is increasing. Many people with diabetes or prediabetes are unaware of their condition, which can lead to serious health complications. Regular health checkups and screenings can help in detecting diabetes and other health conditions in their early stages, which is crucial for effective treatment and prevention of complications.

Research Questions:

1. Which indicators are more correlated towards Diabetes?
2. Which indicator variables has more importance on Diabetes while performing predictive analysis?
3. Which predictive model is the best for your prediction analysis, in terms of time and accuracy?
4. What are the applications to utilize develop this predictive model?

Tools:

- Language: Python
- Module: pandas, NumPy, scikit-learn, matplotlib, seaborn
- Editor: python-notebook

LITERATURE REVIEW

In the paper, "*Exploratory risk prediction of type II diabetes with isolation forests and novel biomarkers"* by Hibba Yousef has discussed about Type II Diabetes mellitus is a rapidly growing serious health concern worldwide. The Disease causes serious life-threatening complications if it is not managed and controlled within time. It is important to identify the T2D patients as soon as possible to save them from the disease complications. In this study, the researcher developed an interpretable machine learning model to identify individuals with a high risk of T2D disease. The ML model worked on the DM-associated biomarkers of oxidative stress (OS), inflammation, and mitochondrial dysfunction (MD). In this model, an Isolated Forest (iForest) was used as an anomaly algorithm detection system. Control group data was used in the model to identify high-risk individuals for T2D. The two iForest models were evaluated through ten-fold cross-validation. The first model worked on the traditional biomarkers (BMI, blood glucose levels (BGL) and triglycerides alone. The second model worked on additional markers such as oxidative stress (OS), inflammation, and mitochondrial dysfunction (MD) including traditional markers. The second model performed better than the first one in all evaluation metrics, particularly for F1 score and recall. The iForest second model finds novel biomarkers such as interleukin-10 (IL-10), 8-isoprostane, humanin (HN), and oxidized glutathione (GSSG). The model is a promising tool to predict and understand the early-stage risk of T2D and provide pharmacological suggestions to manage T2D [1] . In the paper, "*Predictions of diabetes through machine learning models based on the health indicators dataset*" by Xinyi Ren has discussed that Diabetes Mellitus is a widely spread chronic disease in the USA. Individuals lose the ability to control their blood glucose levels and face major complications. This scenario impacts the country's economy and medical expense burden on patients. The objective of the study is to find out the diabetes-associated indicators and develop a diabetes prediction model. The researcher used the original dataset from BRFSS (the Behavioral Risk Factor Surveillance System) and a cleaned dataset from Kaggle for the year 2015. The Kaggle dataset has 253,680 survey responses to CDC (Centers for Disease Control and Prevention)'s and BRFSS with the target variable diabetes and 21 feature variables. The chi-square test was applied to find out the relation between the indicator and DM. Several ML methods were developed for diabetes prediction. The Boost Classifier method was selected with 86.6% accuracy for the testing set. The selected 5 most important features for diabetes prediction were General Health (GenHlth), BMI (Body Mass Index), Age, high blood pressure (HighBP), and high cholesterol (HighChol) variables. The model was less precise due to an imbalance of data. The model can be improved by collecting more diabetic patient data [2] .In this paper, "*Diabetes Prediction using Machine Learning Algorithms*" by Chandrashekar discussed that Machine learning is a tool that utilizes an algorithm to dissect large data for diabetes prediction. It could be helpful for the prognosis of diabetes and lifestyle management to avoid the risk of diabetes. Various diabetes-associated factors such as patient information, lifestyle, and

previous medical history can be analyzed with different machine learning algorithms. These ML algorithms like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), K-Means, K-Nearest Neighbour (KNN), and Naïve Bayes (NB) used to predict the chances of individuals developing diabetes. The PIMA Indian dataset was used to experiment. The algorithm was evaluated for its accuracy and Enthought Canopy software was used to determine accuracy. Prediction result computed prediction results based on precision, recall, and f-measure. delicacy of algorithms. The study concludes that the Naïve Bayes algorithm has higher accuracy to prognose diabetes individuals [3] . In this paper, "*Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique* " by Israt Jahan Kakoly has discussed that Type II diabetes is a globally prevalent disease and it is important to know the associated risk factors with T2D. The research's main focus was to identify or predict the diabetes-associated risk factor using machine learning algorithms. These were namely decision tree, random forest, support vector machine, logistic regression, and KNN. The research applied the Two-fold feature selection techniques i.e. principal component analysis (PCA) and information gain to improve the prediction accuracy. The primary data collected based on the safety procedure described in the Helsinki Declaration, 2013 and 738 records were included in the final analysis. The accuracy level of the result was 82.2% with an AUC of 87.2%. The outcome was significantly good. The research concludes that the inclusion of clinical factors instead of nonclinical factors significantly improves prediction accuracy [4]. In the paper, "*An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study*" by Umm e Laila stated that Diabetes is a chronic condition, increase high blood glucose level. World Health Organization (WHO) conducted a research reveal increase in the number of diabetes cases. Ignorance in diagnosis and treatment of Diabetes cause neuropathy, retinopathy, nephropathy and muscle dystrophy. The prognosis of diabetes and treatment is very important to prevent health risk. Numerous machine learning techniques used for disease prediction but attain accuracy more than 80%. This study aims to increase the accuracy of machine learning ensemble standard algorithms (including AdaBoost, Bagging, and Random Forest) by analyzing the UCI diabetes dataset and comparing their performances. The diabetes data with 17 attributes were collected from the UCI repository which contains different datasets. The dataset utilized 17 attributes reflecting patient and hospital outcomes. The 517 instances including a class attribute used to predict the negative and positive possibility of having diabetes or not. When the ensemble standard algorithms compared during cross validation and found that Random Forest gives the best accuracy, precision, recall and F-measure compared to other ML techniques. The Chi-Square attributes selection approach used to calculate score of individual approach. The attribute polyuria obtained high score (208) and age; itching got lowest attribute (0). The study will be helpful in healthcare system and for people to managing disease. It is necessary to develop advance algorithm methods, gather data, Improve data quality etc. in future.

METHODOLOGY

Algorithms follow to build the models:

Clean the datasets for NA/NULL Values

Many machines learning classifier doesn't work on the dataset which has null values in their samples as it produces error or generate inaccurate prediction, So It's better to clean the dataset before performing any step towards model building. We have applied NumPy library to check and remove the sample consist the null values.

Perform the statistical analysis on each health marker vs Diabetes

We have utilized the Chi-Square test to perform the statistical analysis of each marker vs Diabetes to prove Null Hypothesis False. Null Hypothesis states that any relevance feature vs Diabetes is just luck by chance, there is no direct relevance between under the significance value.

Remove the Outliers from the datasets

We have utilized the box-plot to visualize the outlier of non-binary columns i.e., BMI, Education, MenHlths, Income. Removal of these outlier was performed by utilizing the inter-quartile-range (IQR) to remove the outliers' samples from the datasets. This may change our dataset percentage of Diabetes Vs Non-Diabetes little off balance, but It was the necessary to get the highest accuracy.

Perform the features selection (PCA/Feature Importance)

In our datasets, we have total 21 health markers, which is not directly connect with diabetes, but still they have relation which can be seen in heatmap plot of the datasets. For reducing the dimension of the dataset for better prediction accuracy. We have tried many methods i.e. PCA/ Random Forest Feature Elimination / Extra Tree Classifier Feature Elimination / Recursive Feature Elimination. In most of the methods, we need to specify the dimension which we want to reduce to find the appropriate feature for building the model which can improve the classifier performance.

```
In [34]: clf = RandomForestClassifier(n_estimators=50)

In [35]: clf.fit(X,y)

Out[35]: RandomForestClassifier(n_estimators=50)
         In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
         On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [36]: clf.feature_importances_

Out[36]: array([0.0735559 , 0.0456263 , 0.00614493, 0.16973039, 0.02800026,
                0.01027454, 0.02013958, 0.02317125, 0.0282015 , 0.02203668,
                0.00960417, 0.00759716, 0.01187758, 0.10387265, 0.05268383,
                0.06928449, 0.02641725, 0.02638433, 0.12562125, 0.05731517,
                0.08246076])

In [37]: X.columns[clf.feature_importances_ > 0.05]

Out[37]: Index(['HighBP', 'BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', 'Education',
                'Income'],
               dtype='object')
```

Figure 1 : Random Forest Classifier used for Feature Selection

We have used the Random Forest to build each health marker, feature importance. Which is shown in figure 2. In the next step we have check with many values of threshold for removing the, but if we choose the threshold value below 0.05, it was selecting almost all the features. So we have taken threshold for 0.05, when we did that, It almost selected all the non-binary column (health markers) and few binary column (health markers).

Build the Model, measure converge time, and performance (accuracy)

In machine learning model building, we used to normalize the dataset when the features have many diverse values. In our dataset, all the values categorical values except few such as BMI. Rest of the other non-binary column data is categorical as bin data i.e. Education, MentHlth, GenHlth, Income, PhyHlth, Age. So didn't perform the scaling in our datasets.

Supervised Machine Learning Model:

We have selected fives machine learning classifier i.e. Logistics Regression, Random Forest, Decision Tree Classifier, XGBoost Classifier, Naive Bayes Classifier to build our machine learning model.

Logistic Regression:

Logistic Regression is a statistical Machine Learning algorithm that is used for classification problems. It is based on the concept of probability. It is used when the dependent variable (target) is categorical. It is widely used when the classification problem at hand is binary; true or false, yes or no, etc. Logistics regression uses the sigmoid function to return the probability of a label.

Gaussian Naïve Bayes

Naïve Bayes classifier used the Bayes Theorem for prediction of samples. The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

$$P(c|x) = \frac{P(x|c) * P(c)}{P(x)}$$

- P(c|x) - Posterior Probability of the response (target) variable given the training data inputs
- P(c) - Prior probability of the class (target)
- P(x|c) - Probability of the predictor (x) given the class/target (c)
- P(x) - Prior probability of the predictor (x).

Decision Tree Classifier:

Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions. Decision trees can perform both classification and regression tasks. Decision is kind of an umbrella term, the intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the dataset until you isolate all data points belonging to each class. With this process you're organizing the data in a tree structure. Every time you *ask a* question, you're adding a node to the tree. And the first node is called the root node. The result of *asking a question* splits the dataset based on the value of a feature, and creates new nodes. If you decide to stop the process after a split, the last nodes created are called leaf nodes.

Random Forest Classifier:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

XGBoost Classifier:

XGBoost is an ensemble learning algorithm meaning that it combines the results of many models, called base learners to make a prediction. Just like in Random Forests, XGBoost uses Decision Trees as base learners. XGBoost Library is parallelizable which means the core algorithm can run on clusters of GPUs or even across a network of computers. This makes it feasible to solve ML tasks by training on hundreds of millions of training examples with high performance. Due to Its speed and performance are unparalleled and it consistently outperforms any other algorithms aimed at supervised learning tasks.

Data Analysis

For understanding the diabetes, I have browsed through many data maintaining website i.e. data-world, Kaggle, few hospital website as well which keeps the information of diabetes patients who are facing diabetes as challenges in their life. We have selected the datasets, "Diabetes Health Indicators dataset", was downloaded from Kaggle. This dataset was generated by health-survey named "The Behavioral Risk Factor Surveillance System" (BRFSS) that is collected in 2015 by the "Central of Disease Control" (CDC). Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

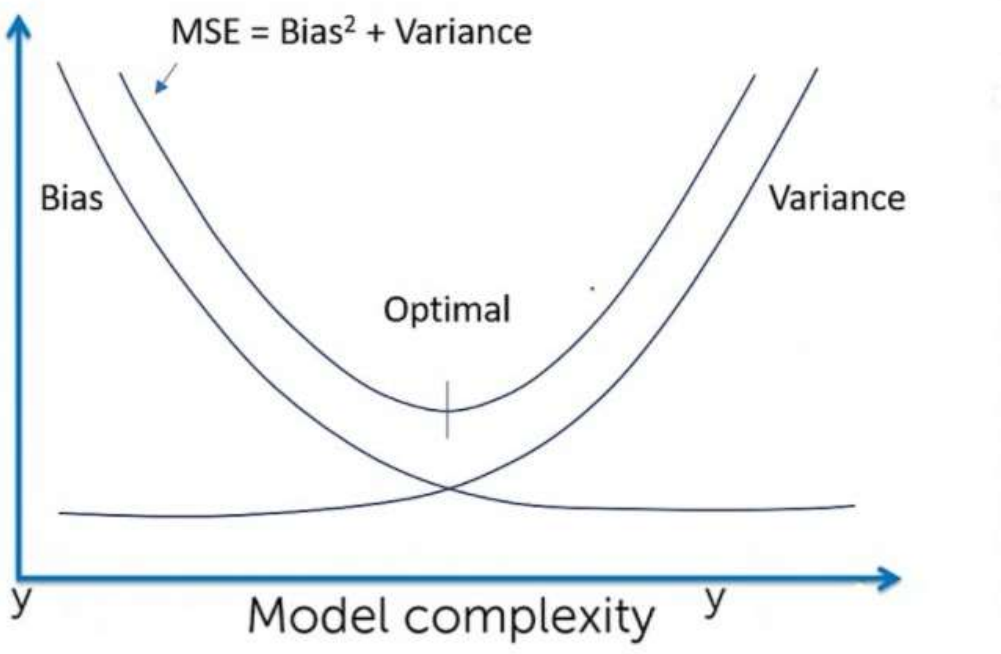
This dataset contains 3 files:

1. diabetes_012_health_indicators_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. This dataset has 21 feature variables and is imbalanced.

2. diabetes_binary_5050split_health_indicators_BRFSS2015.csv: is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced. (Selected for Building Model)
3. diabetes_binary_health_indicators_BRFSS2015.csv: is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

For the selected dataset, we have the written the following important facts:

1. It has many independent variables (20 feature) and 1 dependent variable (diabetes binary)
2. It has many numbers of samples in each kind of dataset files 70692 (Selected CSV File). It is good for predictive model to have sufficient number of samples for training and testing. But
3. It has the features i.e. Diabetes_binary, HighBP, HighChol, CholCheck, BMI, Smokes, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.



CONCLUSION

In this study, we have explored the health marker which impacts a human life to trigger his system to become a Diabetes Person. We have developed diabetes prediction models using machine learning kernels and different dataset cases. The studies found that Age, BMI, family history of diabetes, physical activity, smoking status, waist circumference, systolic blood pressure, fasting plasma glucose, glycated hemoglobin (HbA1c), triglycerides, and total cholesterol were significant predictors of diabetes.

List of References

1. Zhang, Liying, et al. "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study." Scientific reports 10.1 (2020): 4406
2. Laila, Umm E., et al. "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study." Sensors 22.14 (2022): 5247.
3. Shin, Juyoung, et al. "Development of various diabetes prediction models using machine learning techniques." Diabetes & Metabolism Journal 46.4 (2022): 650-657.
4. Rani, KM Jyoti. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305
5. Schneiders, Josiane, et al. "Quality indicators in type 2 diabetes patient care: analysis per care-complexity level." Diabetology & metabolic syndrome 11 (2019): 1-9.
6. Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.
7. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging Risk Factors Collaboration
8. Website : TowardDataScience https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2
9. Kaggle Dataset link : https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2