Automated Diagnosis of Type II diabetes by incorporating machine learning techniques with electronic health records

A MAJOR PROJECT – II THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted by:

SAJAL JAIN

(2K19/SWE/13)

Under the supervision of

Prof. Rahul Katarya

Department of Computer Science & Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY (Formerly Delhi College of Engineering) Bawana Road, Delhi-110042

OCTOBER,2021

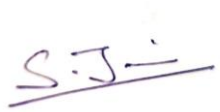
CANDIDATE'S DECLARATION

I Sajal Jain, Roll No. 2K19/CSE/13 student of M. Tech Software Engineering, hereby declare that the major project-II work entitled "Automated Diagnosis of Type II diabetes by incorporating machine learning techniques with electronic health records" Which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfilment of the Master's degree requirement, is authentic and not copied from any source despite appropriate references. Previously, this work did not form the basis for granting any degree, Diploma Associateship, Fellowship, or other specific title or recognition.

Place: Delhi

Sajal Jain

Date: 29-10-21



CERTIFICATE

I hereby certify that the Project Dissertation titled "Automated Diagnosis of Type II diabetes by incorporating machine learning techniques with electronic health records" which is submitted by Sajal Jain, Roll No 2K19/SWE/13, Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University.

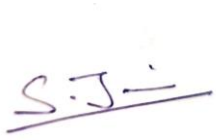
Place: Delhi — Prof. Rahul Katarya

Date: 29-10-21 — SUPERVISOR

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Prof. Rahul Katarya, my major project guide, Professor, Department of Computer Science & Engineering, Delhi Technological University, for the precious advice and encouragement he has given in this major project. It is my delight to record my heartfelt thanks to my esteemed advisor for his valuable feedback and advice that would have not influenced the project as it has. I graciously express my words of gratitude to all other faculty members of this department for providing their invaluable guidance and time whenever requested.

Sajal Jain



Roll No. 2K19/SWE/13

M.Tech (Software engineering)

E-mail: sj31296@gmail.com

Abstract

Diabetes is an irremediable disease from which millions of people are suffering all around the world Diabetes can lead to other life threating problems like kidney failure, eye sight loss etc. Due to Life style of peoples the number of individuals suffering from this disease is increasing everyday Diabetes can even lead to death. Machine learning methodologies can be helpful in detecting diabetes in infancy. In this project we have developed an ensemble model for the detection of ty diabetes. The imbalanced data is balanced using the SMOTE technique. Grid search technique is use finding the optimal values of the hyperparameters. LightGBM and K-NN is ensemble using the soft v classifier. The soft voting classifier will add the prediction probabilities of both the classifiers and pre on the basis of that. The class having greater probability will be predicted by the soft voting class Two datasets are used for analyzing the proposed model. Correlation graph is used for detectin correlated features of the dataset. The proposed model gives accuracy of 90.62 for the pima Indian dia dataset and 94.93 for the Kaggle diabetes dataset. It is found that the proposed model performed bet compared to other state of the art models.

Keywords-Artificial Intelligence, Diabetes Mellitus, Electronic health record, Health Data, Machine Learning

CONTENTS

CANDIDATE'S DECLARATION ii

CERTIFICATE iii

ACKNOWLEDGMENT iv

ABSTRACT v

CONTENTS vi

List of Figures vii

List of Table viii

List of Abbreviations ix

CHAPTER 1 1

INTRODUCTION 1

1.1 Diabetes Overview 1

1.1.1 Types of Diabetes 2

CHAPTER 2 3

Related Work 3

CHAPTER 3 6

METHODOLOGIES 6

3.1 Gridsearch 6

3.2 LightGBM 6

3.3 K-NN 8

CHAPTER 4

PROPOSED WORK 11

4.1 Architecture of the Model 11

4.2 Working of the model 12

CHAPTER 5 14

EXPERIMENT ENVIRONMENT 14

5.1 Datasets 16

5.1.2 Data pre-processing 16

CHAPTER 6 17

Experiment Results 17

6.1 Performance Metrics 18

6.2 Discussion 18

CHAPTER 7 21

Conclusion and Future Work 21

References 22

LIST OF PUBLICATIONS OF THE CANDIDATE'S WORK 25

List of abbreviations

LightGBM: Light Gradient Boosting machine

GB DT: Gradient Boosting Decision Tree

K-NN: K Nearest neighbour

DM: Diabetes Mellitus

RNN: Recurrent neural network

SVM: Support Vector machine

List of Figures

Fig 1 Representation of K nearest Data points……07

Fig 2 Leaf and level-wise Strategy……09

Fig 3 Structure of the Proposed Model……11

Fig 4 Overall overview of the dataset…… 14

Fig 5 Correlation between the features of Pima Indian Diabetes dataset…… 15

Fig 6 Heat map of the Kaggle diabetes dataset…… 15

Fig 7 Performance of the model……19

Fig 8 comparison in terms of accuracy…… 19

Fig 9 comparison in terms of f1-score……20

Chapter 1

Introduction

1.1 Diabetes overview

The biotechnology area is going through a lot of changes and enormous amount of data takes the field towards the area of big data and artificial intelligence. A lot of data is coming from various provenance and in different forms like images and text[1]. So it requires methodologies to analyze the data and make meaningful conclusions using that data. One of the application of using these kind of technologies is to predict the life threating diseases in early stages so that the lives and resources can be saved. Diabetes metilus is one of those disease which can also leads to other life threating problems[2]. All around the world many individuals are suffering from this malign disease. One of the symptom of disease is the rise in glucose level for a prolonged period of time. Insulin is secreted by the pancreas. When one body is unable to produce insulin in sufficient amount or is unable to use the insulin produced by the body can lead to diabetes[3]. There are various risk factors which affect the diabetes like pregnancy, age etc. Type II diabetes is the most habitual of all diabetes types. Type II diabetes occurred when there is not enough insulin produced by the body or one's body is unable to utilize it.

Adult-onset diabetes is another name for type II diabetes. Certain measures can be taken if diabetes is prognosticating in early stages which eventually leads to saving lives and resources of the individuals. The lab tests which are used for the detection of diabetes are very tedious and time taking so there is a need to construct an self-operating and faster approach to prognosticate this chronic disease. Artificial intelligence can be used for this purpose[4]. Artificial intelligence can be used to prognosticate diabetes in early stages by making use of

the medical records of the patients. Several Machine learning methodologies had been proposed for the early prognosticate of diabetes. Machine learning can be utilized to prognosticate diabetes by using the medical records.

1.2 Types of Diabetes

Diabetes is an irremediable disease which eventuate in body when body is unable to produce glucose in sufficient amount or unable to make use of it. Due to which it can move sugar to the blood from the cells of the body. Diabetes can be classified into four types[5]

1)Type I diabetes – In type I diabetes the pancreatic cells that produce insulin are destroyed somehow.

2)Type II diabetes-In type II diabetes the body produces insulin but it can utilize it. There are many factors which can lead to type II diabetes like lack of exercise, lifestyle etc.

3)Type III diabetes- Type III diabetes is the association of Type I and II diabetes which leads to Alzheimer's. Alzheimer occur due to the resistance in the brain for insulin.

4)Gestational Diabetes- It is found only in pregnant woman's body. When a woman is pregnant sometimes some hormones are generated which blocks the insulin producers.

Related Work

In this Section, we have discussed the work done previously in this area. By doing proper research and analysis we have found the shortcomings of the previous works and tried to improve it by the proposing a different predicting model.

In [6] Author have used different machine learning algorithms for the prognosticate of Type II diabetes. The proposed model is developed by incorporating two machine learning methodologies which are Particle swarm optimization(PSO) and Fuzzy Clustering Means(FCM). The proposed model(PSO-FCM) performed better than other proposed models but authors did not used sufficient performance metrics for comparison with other proposed models.

Referring to [7],Authors have incorporated innumerable machine learning Techniques for the prediction of diabetes. The authors have used six machine learning methodologies for predicting the type II diabetes. These 6 methodologies are the basic machine learning methodologies which are availed in the domain of machine learning. Authors have used 10-fold cross validation technique for cross validation. For the comparison they used various performance metrics to evaluating the performance of the methodologies. They concluded that random forest performed better than the other machine learning methodologies.

In the article [5], Authors have used Extreme gradient boosting(XGBoost) for the diagnosis of type II diabetes and weighed against several other machine learning models like Support Vector Machine (SVM), Random Forest(RF) and K-nearest neighbour. For evaluating the performance of the model authors have used a dataset which is collected through a

questionnaires. For cross validation they used 20-fold cross validation technique. For the comparison of the models authors have used different performance indicators.

In reference [8], The authors have proposed an ensemble model for the classification of the type II diabetes. The ensemble model is developed by ensembling three machine learning methodologies which are random forest, Logistic regression and Naïve Bayes. The proposed model is compared with several other models. Authors have used the proposed model for the prediction of breast cancer also. The authors concluded that the proposed model performed better than various state of the art models.

Referring to [9], Authors have proposed a model which utilizes the auto stack encoders and deep neural networks. The auto stack encoders are used for extracting the features form the dataset. Then the dataset is classified by employing the softmax layer. The Backpropagation strategy is used for fine tuning of the neural network. The authors have used various performance indicators for evaluating the capability of the model. The authors concluded that the proposed model outmatch than several other proposed models.

In Article [10],The authors have used various ensemble models for the prognosticate of type II diabetes. The authors have used several techniques for the pre-processing off the dataset. The outliers are removed in the data pre-processing. For cross validation the authors have used K-fold cross validation technique. The proposed framework utilizes various state of the art models. In addition, the authors have also utilized multilayer perceptron(MLP) for the prediction of diabetes. The authors ensemble all the models for the diagnosis of type II diabetes. The authors concluded that the ensembling of XGBoost and Adaboost performed better than all the other models.

Referring to [11],Authors have Used Various machine learning methodologies for the prediction of diabetes. The authors have employed six machine learning methodologies for predicting the type II diabetes. These 6 methodologies are the basic machine learning methodologies which are employed in the domain of machine learning. Authors have used 10-fold cross validation methodology for cross validation. For the comparison they used various performance metrics to evaluating the performance of the methodologies. They concluded that random forest performed better than the other machine learning methodologies.

Chapter 3

Methodologies

1.Grid search

Grid search is a technique that is utilized for finding the optimal values of the hyperparameters. Hyperparameters whose values are necessary for managing the training process of a machine learning model. It is mandatory to discover the optimal values of these parameters. For discovering the optimal values one medium is to do a comprehensive exploration of all the feasible solutions and pick the one which has the minimum fault f. The searching of every possible value is not feasible but it is possible to search for the optimal values from a large set of results[12]. when a grid is used to represent the possible values of the hyperparameters the technique is known as Grid search. Let the dimension of the grid is p and for every dimension, the number of results is q then there will be q^p points that should be explored for generating the optimal values of the hyperparameters. To make the Grid search technique optimal constraints should be defined to limit the search operation. Despite being an easy technique it does not perform well when there is a lot of hyperparameters for a model[13].

2.K-Nearest Neighbour

K-Nearest neighbour is a machine learning methodology, which despite being elementary it brings out remarkable results[7]. This methodology can be utilized for various kind of problems. In KNN, the K nearest data points are discovered and then the data point is assigned based on the dominant class found in the k nearest data points. The best advantage of K-NN is that if applying a model on some training data exasperating K-NN can be utilized to find the optimal solution[14]. So as result K-NN helps in recognizing beneficial patterns and used in diverse problem across the artificial intelligence domain. The whole procedure of K-NN comprises two parts first part is the training of the model and the second part is forecasting using the test data. The most important thing in the training phase is finding the optimal value of k. The selection of k data points is done in a procedure[15]. This procedure involves finding the distance from all the data points then arranging them in ascending order and select the first k data points. For calculating the distance between the data points the KNN utilizes the Euclidian formula which is represented by eq (1)[16].

$$D(p,q) = \sum_{m=1}^{n} \sqrt{(x_m - y_m)^2} \qquad (1)$$

The forecasting for the test data is based on the majority law which is very homogenous to the Bayesian rule.

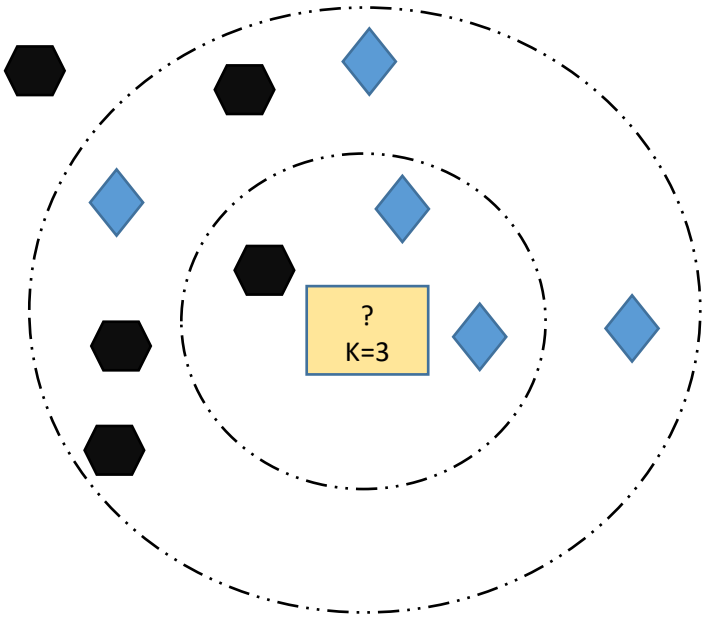


Fig 1 Representation of K nearest Data points

There are three demanding tasks in the K-NN algorithm: determining the best value for k, searching for the k nearest neighbours and defining predicates for the classification. In Fig 1 The pink data point is the test data for which forecasting is to done and the value of k is 3. So as the value of k is 3 the nearest data points are represented in the inner circle. By considering the Majority rule, the data points having blue colour are in majority so the test data point will be allocated the class of the data points represented by diamonds[17]. Despite being all the advantages associated with K-NN there is a crucial drawback associated with it which is its Mediocre performance when there is an escalation in the number of variables or the data is expanded[18].

3.LightGBM

There are various reasons behind the popularity of the Gradient boosting decision tree (GBDT) like efficiency and performance for classification problems. The standard enforcement of GBDT inspects each data point for every single feature which is not feasible[19]. There is a simple approach for the above problem that is to reduce the number of data points and features which is also not feasible. So there is a need for an implementation that tells us how to perform sampling on the data[20]. Two techniques can be used for the above process .one is Gradient-based one side sampling(GOSS), In GOSS as there are no weights associated with data points so gradients are taken into consideration. The data points which have a large value of the gradient have a greater impact on the results. So samples having smaller gradients are deleted irregularly. The GOSS technique over performed as compared to arbitrary sampling.

Another one is Exclusive feature bundling (EFB), In EFB we can reduce the number of features by not losing any data but the condition is that the feature space should be sparse. Generally, the feature space of real-world applications is large and sparse. By incorporating the above two techniques (GOSS and EFB) LightGBM technique is created. As it is an ensemble methodology that consists of various weak learning models (Decision trees). As the other algorithms work on the level-wise scheme LightGBM follows the leaf wise scheme. Both the schemes are shown in fig 2.

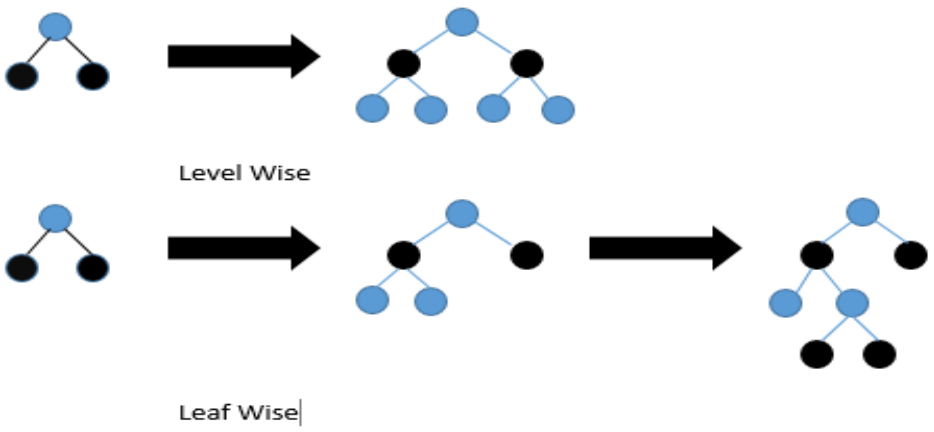


Fig 2 Leaf and level-wise Strategy

There is one problem that arises in the leaf wise strategy that is the overfitting of the model. The solution to the above problem is that the depth should be predefined.

Advantages of LightGBM

- The Potential of LightGBM in terms of accuracy and efficiency is better in contrast to other models.
- The Consumption of memory is lower.
- Parallel learning is also present.
- Higher training speed as compared to other models.
- LightGBM performs very well with datasets of all sizes

Chapter 4

Proposed Work

In this section, we will discuss the proposed hybrid model (GBM-KNN) that incorporates the LightGBM and K-NN methodologies for the premature discovery of Type II diabetes. The datasets are collected from the Kaggle machine learning repository. The feature engineering techniques are applied to the datasets. After the data pre-processing is done the dataset is split into two parts. These parts are known as training and testing data.

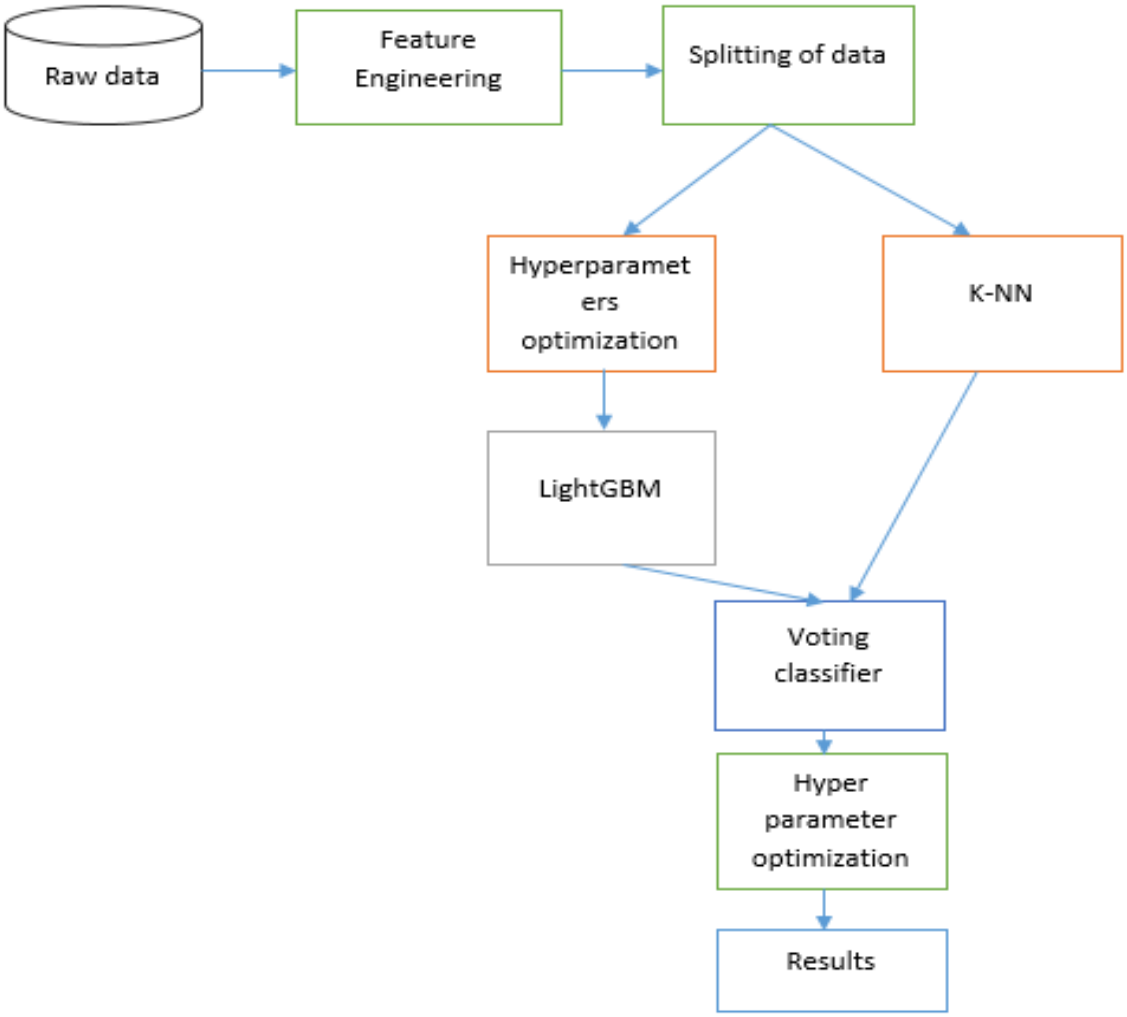


Fig 3 Structure of the Proposed Model

In the proposed model for hyperparameters optimization, the Gridsearch technique is used. For applying the Gridsearch technique GridsearchCV method is used. The randomized search is also applied for the comparison with the grid search. After the hyperparameters optimization, LightGBM is trained on the training data with the values of the parameters obtained by the Gridsearch method. The trained data is also used to train the K-NN classifier. After both, the classifier is trained a soft voting classifier is used to ensemble both the classifiers.

4.1 Working of Proposed model

- The first step requires the extraction of the datasets. The datasets are extracted from the Kaggle machine learning repository for the early discovery of diabetes.
- Data pre-processing methodologies should be employed to the raw data so that the raw data can be modified into a meaningful data format
- Data cleaning methodology is used in which we remove the unwanted zeroes and null values present in the dataset.
- Before training a model the data should be normalized.
- Normalization is a technique in which the range of data of an attribute is decreased.
- To check normalization of data different test is used In our experiment we have used Shapiro Wilkinson's test.
- Better and efficient results are obtained with normalized data.
- After normalization correlations are also checked in the datasets by utilizing a correlation graph.
- Data is now ready for the model to be trained.
- After data pre-processing the data is divided into two parts training and testing
- LightGBM is the first technique used in the creation of the proposed model.
- To stave off overfitting of the model maximum depth should be defined
- For finding the optimal values of the parameters grid search method is used
- Another classifier is taken for the development of the hybrid model which is K-NN
- For combining both the classifiers a soft voting classifier is used which will combine both of the models

- Again hyperparameters optimization is done for the soft voting classifier so that the optimal values of the parameters can be known.
- The obtained outcomes are then weighed up against other state of the art methods for the detection of diabetes.
- The advantage of using LightGBM is that is take less time and less storage for the premature discovery of diabetes

Chapter 5

5.1 Datasets

5.1.1 Pima Indian diabetes dataset

Pima Indian diabetes dataset contains the medical records of the females who are above 21 years old and belongs to the Pima community living near the Glia river in Arizona. The dataset contains 8 features and 768 medical records of the patients. The 8 features are all

related to diabetes like pregnancy, age and others. Only one feature is a dependent feature.

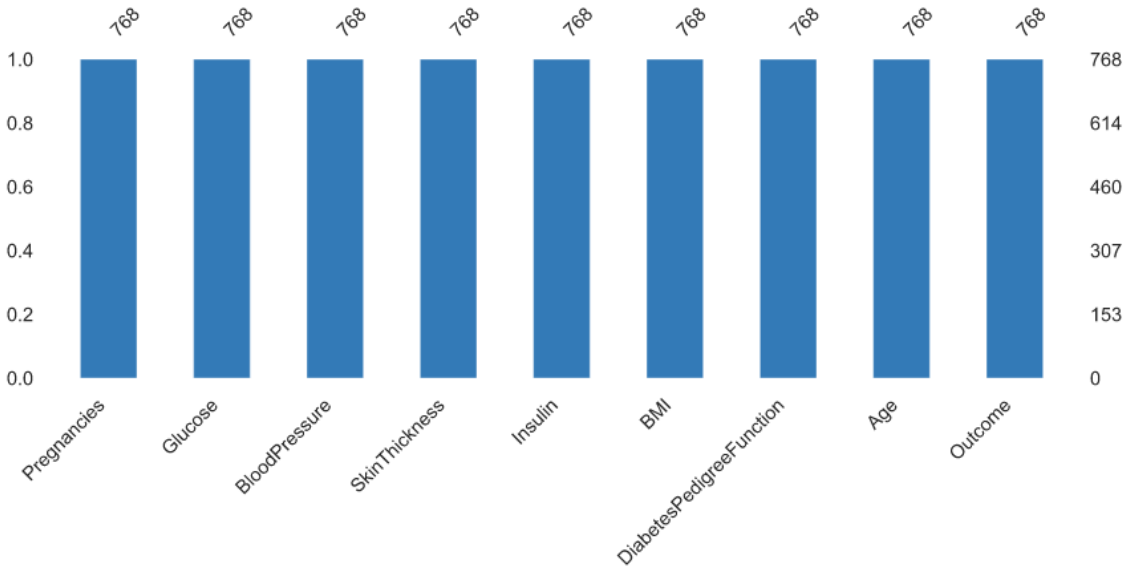


Fig 4 Overall overview of the dataset

Figure 4 and 5 shows the overall overview and correlations between the features of the datasets.

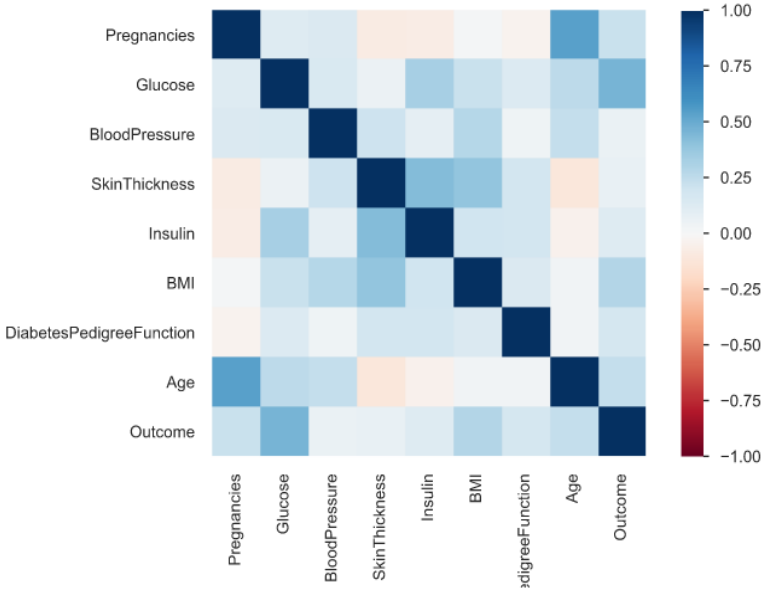


Fig 5 Correlation between the features of Pima Indian Diabetes dataset

5.1.2 Kaggle diabetes dataset

Kaggle diabetes dataset contains the medical records of the 952 patients. In addition to 8 features of the above datasets it contains 10 more features related to diabetes like smoking sleeping habit. The medical records are collected through an online feedback system. Figure 6 shows the heat map of the dataset.

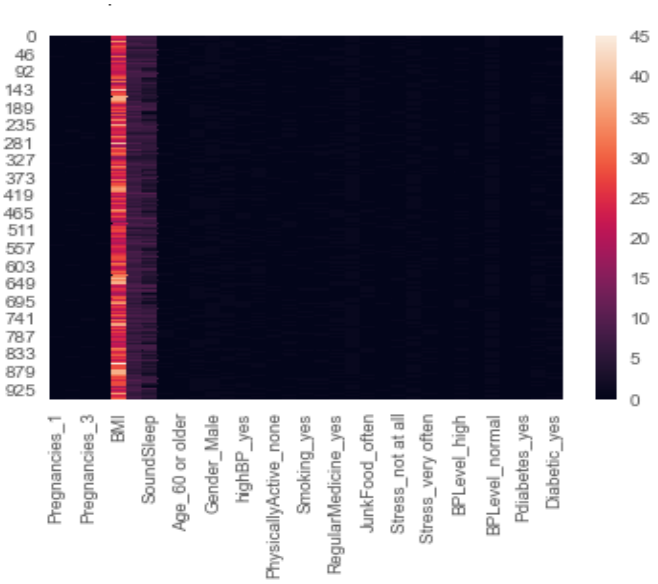


Fig 6 Heat map of the Kaggle diabetes dataset

5.2 Data pre processing

Both the datasets contain null and zeroes present in every feature. The null and zeroes of the dataset are replaced with the median value of the feature. The normalization in the dataset is checked using the Shapiro Wilson test. For detecting outlier's violin graphs are used.

Chapter 6

EXPERIMENT RESULTS

6.1 Evaluation metrics

Six performance measures are utilized to analyze the model's performance which are “accuracy, precision, recall, Roc-Auc and f1-score”. These performance metrics are also used for comparing the proposed model with the other state of the art models. These metrics will help us to judge whether the model is efficiently predicting the adult onset diabetes or not.

Accuracy- Accuracy is the ratio of the instances where the model predicts the type 2 diabetes correctly to the total number of instances

Precision- Precision is the ratio of the number of instances where the model correctly predicts the diabetes for diabetic patients from the instances where the patient actually has the diabetes.

Recall- Recall is the ratio of the number of instances where the model correctly predicts for a non-diabetic patient from the total instances of the non-diabetic patients

F1-Score- The harmonic mean of recall and precision is known as f1-score

Roc-Auc Score- This precision metric will tell us how efficient our model is in distinguishing between the classes

Confusion Matrix- The confusion matric helps to estimate to performance of the model. It contains four things true positive, true negative, false positive and false negative. With the help of these values we can also estimate the values of other performance indicators. True positive will tells the cases where the model correctly discovered diabetes for the diabetic patients. True negative is the number of cases where the model correctly identifies the non-diabetic patient. False negative is the proportion of cases where the model fallacious identifies a non-diabetic patient and false positive is the number of cases where the model fallacious identifies the diabetic patient.

6.2 Discussion

The model is applied on the two datasets all the results are shown in the fig 5. Profuse machine learning methodologies are employed for the prognosticate of type 2 diabetes. The

fig 4 shows the discriminant analysis of the dataset. The linear discriminant analysis is a methodology which helps to deplete the number of features while preserving the data. The linear discriminant analysis will help to look for the attributes that can help to distinguish between the classes and objects. For the training of the model two datasets are taken all the experiments are performed on the Jupiter notebook the model performance is evaluated by taking five performance metrics into consideration the f1-score is chosen as the basis of comparison of the models. The f1-score can tell us the efficiency of the model for the prognosticate of adult onset diabetes. The proposed model achieved 94.92 accuracy for the dataset A and 90.6 accuracy for the dataset B. The value of f1-score for dataset A is 93.73 and 86.47 for dataset B.

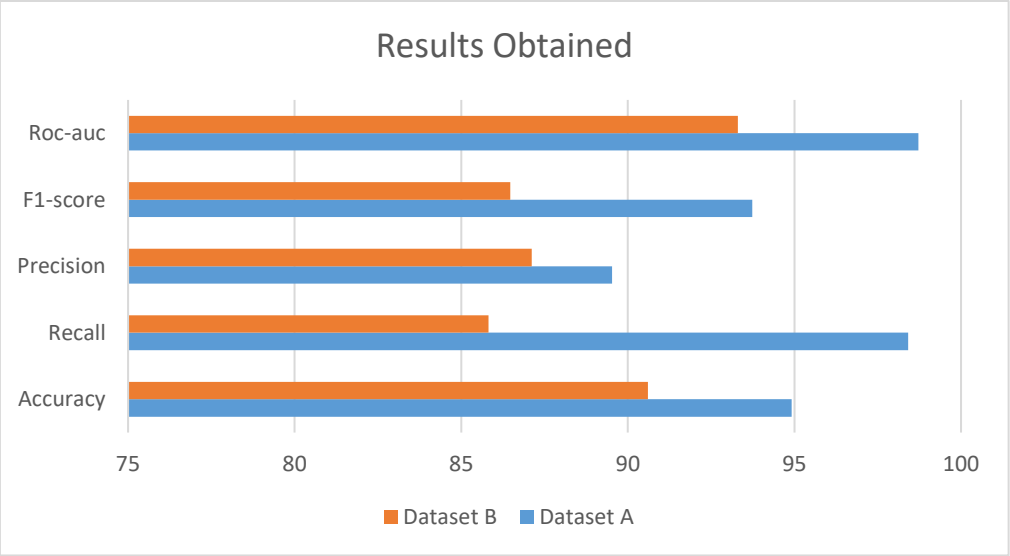


Fig 7 Performance of the model

The model is put up against other state of the art models which is shown in fig 6 and 7.

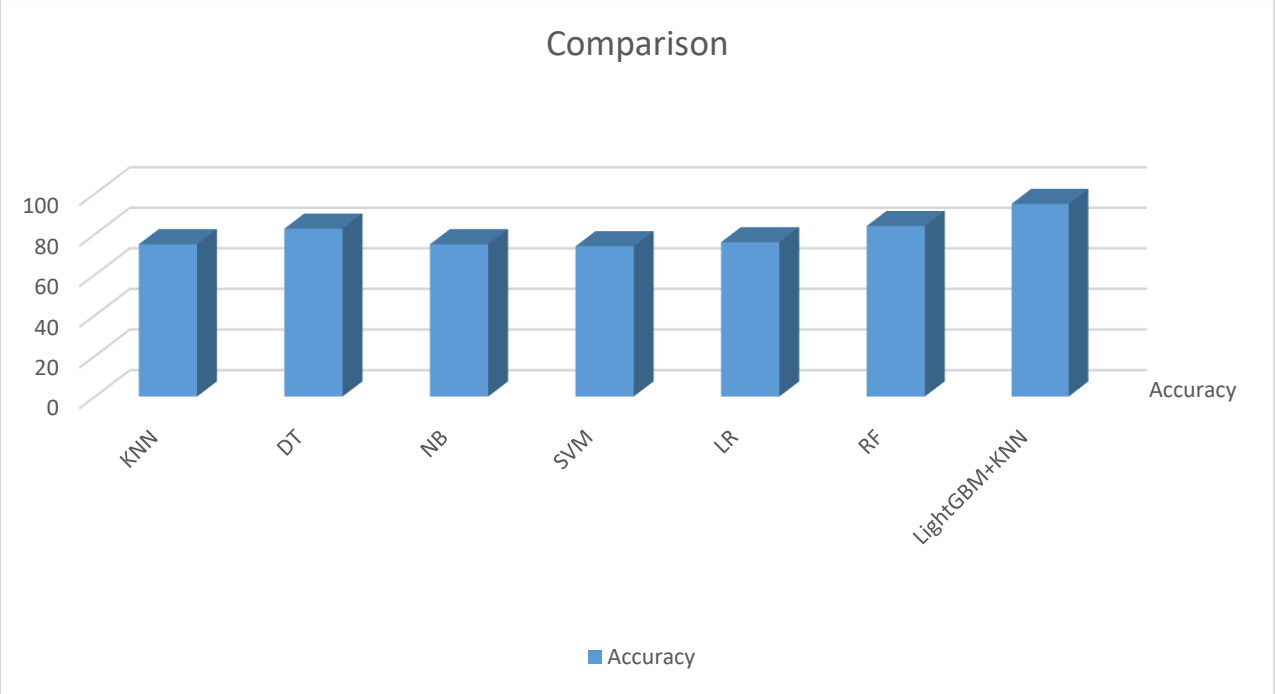


Fig 8 comparison in terms of accuracy

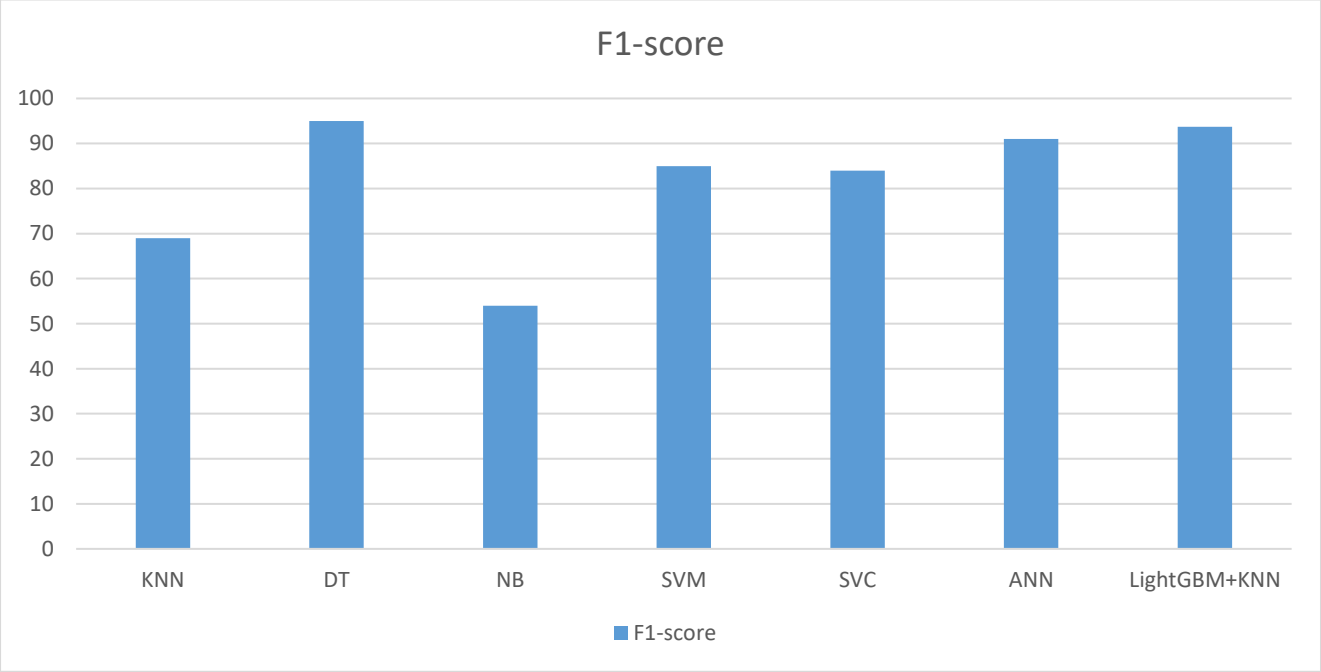


Fig 9 comparison in terms of f1-score

Chapter 7

Conclusion and Future Work

In the modern world many individuals are getting affected from the diabetes. The number of individuals suffering from this are increasing exponentially. There are many health related consequences associated with the diabetes which can eventually led to death. In this study we have incorporated many machine learning techniques for the prognosticate of type 2 diabetes. We have proposed a hybrid model which utilizes LightGBM and K-NN for the prognosticate of type 2 diabetes. The proposed model can predict the type 2 diabetes by utilizing electronic health records of the patients. Human intervention is also lessened by the implementation of the proposed model. We have observed that proposed model performs wonderfully and out performs other models for the prognosticate of type 2 diabetes. In the future we will try to incorporate more techniques for the prognosticate of type 2 diabetes.

Chapter 7

References

[1] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, “Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression : a comparison of four data mining approaches,” pp. 1–13, 2020.

[2] N. Li, J. Tuo, Y. Wang, and M. Wang, “Prediction of blood glucose concentration for type 1 diabetes based on echo state networks embedded with incremental learning,” *Neurocomputing*, vol. 378, pp. 248–259, 2020.

[3] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, “Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review,” *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2020.

[4] O. Ben-Assuli, T. Heart, N. Shlomo, and R. Klempfner, “Bringing big data analytics closer to practice: A methodological explanation and demonstration of classification algorithms,” *Heal. Policy Technol.*, vol. 8, no. 1, pp. 7–13, 2019.

[5] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, “Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model,” *Healthcare*, vol. 8, no. 3, p. 247, 2020.

[6] J. B. Raja and S. C. Pandian, “PSO-FCM based data mining model to predict diabetic disease,” *Comput. Methods Programs Biomed.*, vol. 196, 2020.

[7] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 706–716, 2020.

[8] S. Kumari, D. Kumar, and M. Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,” *Int. J. Cogn. Comput. Eng.*, vol. 2, no. November 2020, pp. 40–46, 2021.

[9] K. Kannadasan, D. R. Edla, and V. Kuppili, “Type 2 diabetes data classification using stacked autoencoders in deep neural networks,” *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 4, pp. 530–535, 2019.

[10] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes prediction using ensembling of different machine learning classifiers,” *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[11] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–14, 2020.

[12] G. Li, W. Wang, W. Zhang, Z. Wang, H. Tu, and W. You, “Grid search based multi-population particle swarm optimization algorithm for multimodal multi-objective optimization,” *Swarm Evol. Comput.*, vol. 62, p. 100843, Apr. 2021.

[13] P. Liashchynskyi, “Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS,” no. September, 2020.

[14] S. Zhang, “Challenges in KNN Classification,” *IEEE Trans. Knowl. Data Eng.*, pp. 1–13, 2021.

[15] W. Cherif, “Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis,” *Procedia Comput. Sci.*, vol. 127, pp. 293–299, 2018.

[16] J. Huang, Y. Wei, J. Yi, and M. Liu, “An improved knn based on class contribution and feature weighting,” *Proc. - 10th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2018*, vol. 2018-Janua, pp. 313–316, 2018.

[17] R. Katarya and S. Jain, “Comparison of different machine learning models for diabetes detection,” *Proc. 2020 IEEE Int. Conf. Adv. Dev. Electr. Electron. Eng. ICADEE 2020*, pp. 0–4, 2020.

[18] C. Li *et al.*, “Using the K-Nearest Neighbor Algorithm for the Classification of Lymph Node Metastasis in Gastric Cancer,” vol. 2012, 2012.

[19] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, “Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data,” *Agric. Water Manag.*, vol. 225, p. 105758, Nov. 2019.

[20] F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, “Improving the use of big data analytics within electronic health records: A case study baseD OpenEHR,” *Procedia Comput. Sci.*, vol. 127, pp. 60–68, 2018.

List of publications

[1] Katarya, Rahul and Jain, Sajal "Exploration of big data analysis in the Healthcare Sector, International Conference on Computer, Communication and Signal Processing" (can be accessed through https://ieeexplore.ieee.org/document/9315192)

[2] Katarya, Rahul and Jain, Sajal Comparison of different Machine Learning Models for Diabetes Prediction, International Conference On Advances and Development in Electrical and Electronics Engineering (can be accessed through https://ieeexplore.ieee.org/document/9368899)

[3] Katarya, Rahul and Jain, Sajal An Ensemble Model based on LightGBM and K-NN for early Prediction of Type II diabetes, Interdisciplinary Sciences: Computational Life Sciences