Prediction of Type- 2 Diabetes using the LGBM Classifier Methods and Techniques

B. Shamreen Ahamed[1], Dr. Meenakshi Sumeet Arya[2]

[1]Department of Computer Science and Engineering,Research Scholar,SRM Institute of Science and Technology, Vadapalani Campus, Chennai
[2]Department of Computer Science and Engineering,Faculty of Engineering and Technology,SRM Institute of Science and Technology, Vadapalani Campus, Chennai
[1]sb3319@srmist.edu.in,[2] raina.arya@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: In our day to day life we come across a lot of diseases and one of the most commonly heard Auto-immune disease is Diabetes Mellitus. Diabetes is a non-communicable disease, that has affected the lives of many people due to its effects and medications available as of today. There are a number of medical facilities available in the healthcare industry, however a specific computational technique to predict and detect diabetes is not available. Inorder to overcome the problem, a prediction based model needs to be developed. To develop a model, a dataset containing patient details is required along with the necessary attributes needed for testing the presence of Diabetes on the patient. The process is divided into training, validation and testing. There are many Machine learning algorithms available for the study. Some of them include XGB Classifier, Logistic Regression, Gradient Boosting Classifier, Decision Tree, Extra Trees Classifier and Random Forest and LGBM Classifier Algorithm. Out of all the above algorithms, the LGBM Classifier Algorithm is considered to give the most accurate results. The LGBM is a Light gradient Boosting Algorithm which can be implemented using classifiers. The PIMA Indian Dataset is used in this study for the comparison of the different algorithms mentioned above and an accuracy of 95.20% is obtained using the LGBM Classifier Algorithm. Therefore the LGBM classifiers can be used to develop a data model for detecting and predicting diabetes.

Keywords: Diabetes Mellitus, Prediction Model, Training Data, Testing Data, LGBM Algorithm, Accuracy.

1. Introduction

Diabetes Mellitus is one of the Auto Immune Disease that may be caused by a number of factors such as hereditary, environmental conditions, food intake etc. that is growing fast among many other disease in today's era. Diabetes Mellitus can be of two types, Type-1 which commonly affects children and adolescent and Type-2 Diabetes mostly appears in aged people[3]. Diabetes as such is caused by the change in the blood-sugar level(glucose) in the body. It is caused when the maltose level present in the blood cannot be controlled by the body. The insulin content produced by the body has no response from the cells of the body[7]. This is currently being treated by using insulin injections in certain cases and by continuously monitoring through diet and exercise in some cases depending on the intensity of the Glucose level. Some of the early symptoms of diabetes is frequent thirst, itch skin, headache, frequent urination, tiredness etc. The people affected and unaffected are not considering the serious causes and effects of the disease and how it can prevented. Even if prevention is not possible in certain people, some precautionary measures can be taken to help an individual affected by diabetes[3]. There are many advanced technologies in the healthcare industry, however there is no particular cure that is immediate for the diabetic disease[5]. This causes many diverse effects in the health of human. Many research has been conducted for identifying the root cause and specific medications for the particular patient. Big Data Analytics, Data Mining, Deep Learning are some of the genres that are used[5].

Big Data Analysis plays a vital role in obtaining data from large amount of information. It is used as a tool to design the data collected and analyze it. Data Mining is used to retrieve data from databases and store as dataset that can be used later for analysis. The other tools that are used for health related concerns are statistical tools, linear regression, multiple regression, clustering, regression techniques etc[32]. By using these technologies, many datasets are created and the relevant information of the patient affected by the disease is given and the output is predicted. This improves the facilities in the medical industry and results in lesser affected rate of the disease[33].

A number of algorithms are developed using the classification techniques such as Regression, Decision Trees(DT) and Support Vector Machines(SVM)[12]. Many association algorithms can also be used for the same. However, as of today's advancement no such data model is developed for predicting and detecting diabetes as a whole. If such a data model is created it will become easier to overcome the diabetes disease in the future[9].

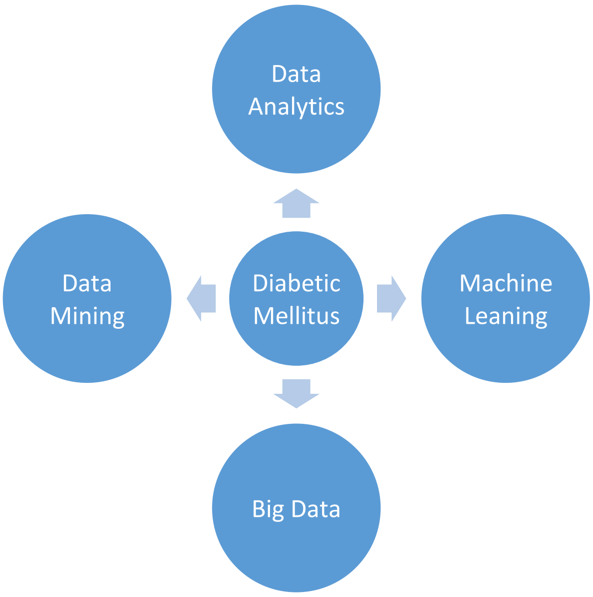
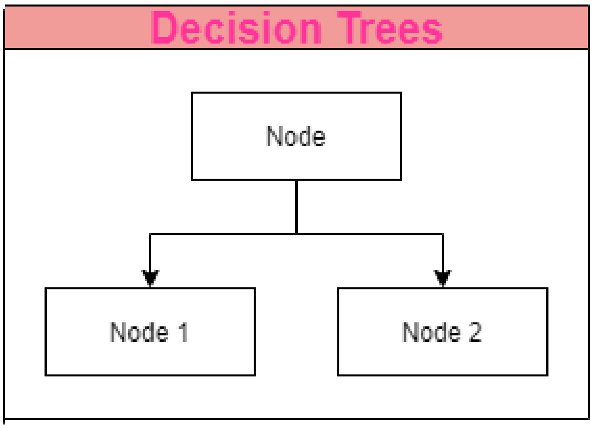


Fig 1.1 Areas for Diabetes Mellitus

The hospitals consists of data necessary about the patient. These data can either be structured or unstructured. The most commonly used data is the raw data that has been received directly from the patient. The disadvantage of raw data is that it increases the steps of analyzing and categorizing data[11]. Based on the variants in which the data is characterized, the results obtained varies.

To overcome this problem, the data is selected and converted into data frames and from that it is further proceeded. This data is then separated as training data and testing data. The training data consists of datasets that already has some knowledge, whereas the data under testing are newly given the knowledge[25]. A validation step is carried out in between the training and testing procedures to give a clearer picture of data that can be used for the study. By doing the above procedures, Channeling of data is done[1].

Some of the algorithms invloved in the study include Decision Tree, Extra Trees Classifier, Logistic Regression, XGB Classifier, Random Forest, Gradient Boosting Classifier and LGBM. “The LGBM Algorithm depends on decision tree algorithms, it divides the tree leaf wise with the best fit though other boosting algorithms split the tree profundity wise or level wise as opposed to leaf-wise” [19].



In the above diagram, each leaf is considered as a type of data and its branches are considered as the different sub groups to each data involved[14]. Therefore, when developing on a leaf from the already existing one, the accuracy is improved and this brings more improved advanced approach to retrieve information and process it using the different algorithms involved. It occurs very fast and henceforth obtains the name “Light”[15].

In this paper the comparison of the research being carried out for detecting and predicting diabetes using the various algorithms are compared and studied.

The paper is categorized as follows:

Section 1 gives the fundamental information about the Diabetes Mellitus Disease, Section 2 provides the ways in which the different techniques can be used, Section 3 provides details about the various studies and work observed for the diabetes disease, Section 4 focuses on the different algorithm used and the prediction methods for the same. In section 5, the results are discussed and obtained. The conclusion and future work are given in Section 6

2. Related work

In the medical field, many ML algorithms are used for detecting and predicting diseases. Diabetes is one such disease that uses ML techniques to obtain the most accurate solutions for Diabetes.

Mercaldo et al.[16] in their study, used 6 different classifier algorithms which included Multilayer Perceptron algorithm, JRip algorithm, Hoeffding Tree algorithm, J48 algorithm, BayesNet algorithm and Random Forest algorithm. The authors in this study used the PIMA Indian Dataset. [16]. The authors have taken two commonly used algorithms namely Best Initial and Greedy Stepwise, to check the differential characteristics that help in defining the concept classification used. Four characteristics were taken specifically for their research which are age, BMI, plasma glucose concentration and diabetes pedigree function . “A ten – fold - cross veradication is pragmatic to the dataset. The classifiers were allied and were thru liable on the value of the recall, accuracy and the F-Measure. The result showed the accuracy in the value that was equivalent to 0.757, F-measure equals to 0.759 and recollection equals to 0.762. From the Hoeffding Tree algorithm we can conclude that it formed the highest presentation linked to others”[16].

Geshwaree et al.[8] used the concept of Glucose Prediction to monitor the diabetes disease. The dataset used is taken from 442 patients manually and determined based on some algorithms. The algorithms used are based on Auto-regressive models (ARX), Blood Pressure (BP), Total Cholesterol (TC), Low-density Lipoprotein (LDL) and High-density Lipoprotein (HDL). The main aim of this study was to present comprehensive critical forecasting review on the prediction models of the recent glucose detection and to obtain the best fit based on wireless body area network system.[8]

Holden et al.[18], introduced new tools for classification and monitoring of auto immune diseases in general. They identified the concept of actionable markers that work as biological metrics that can inform clinical practice. These biomarkers are also used for diagnosing a particular disease. Some of the analytical tools used are automated multivariate estimation (FLAME), density based merging (DBM) and density normalized events (SPADE). These are used for computation and specialization[18].

Sisodia et al. [27] used Decision Tree algorithm, SVM algorithm and Naive Bayes classifiers algorithm for the prediction of diabetes. The main step was to recognize the classifier having the highest accuracy. The dataset used is PIMA Indian Dataset. According to the author, “the 10-folds cross-validation partition was done. The performance was evaluated using the measures of the recall, accuracy, the precision and the F-measure. The Naive Bayes obtained highest accuracy, measuring 76.30%”[27].

Sajal et al.[24] has done a comparative study on the different Machine Learning Algorithms that can be used foe diabetes detection. The various Machine Learning used by them are Support Vector Machine(SVM), Gradient Boosting, Decision Tree, K-Nearest Neighbour(KNN), Logistic Regression and Random Forest. The dataset used is PIMA Indian Dataset. The data is divided in training data(70%) and Testing Data(30%). The authors have used the concept of Python Data Manipulation Tool and have obtained the accuracy for the diabetic detection. Random Forest has the highest accuracy rate of 83%, while the other algorithms have an accuracy of KNN 74.60%, DT 81.6%, NB 73.6%, SVM 73.7% and LR 75.5%. This can be further improved by using the ensemble machine learning methods[24].

Yuvaraj et. al. [34] used 3 machine learning algorithms for their study which are Naive Bayes Algorithm, Decision Tree Algorithm and Random Forest Algorithm.The dataset used was PIMA Indian Dataset. The authors divided the data into training data and testing data. The data was not pre-processed, however the Information Gain method was used for selecting the feature.In total 13 attributes were used out of which 8 were considered active. The algorithm gave an output of 94% using the Random Forest Technique[34].

Olaniyi et al. [23] implemented the concept of Back Propagation Algorithm for the Multilayer Feed-Forward Neural Network. They used the PIMA Indian Dataset for their study. The dataset was initially normalized and then processing of data was carried out. The data was divided into 500 samples for training and 268 samples for testing. An accuracy of 82% was obtained[23].

Soltani et al. [29] used the “Probabilistic Neural Network (PNN) to predict diabetes disease”. The algorithm was applied to the “PIMA Indian dataset”. The author did not apply the pre-processing technique. However, “the dataset is alienated into 10% for the setting customary and 90% for the training set. The projected technique attained exactness of 81.49% for testing and 89.56%, for taxing data”[29].

John Martinsson et al.[14] in their research presented a neural network that is prototypical and is prophesied based on blood glucose levels simplifying the estimate level of vagueness during the prediction procedure. The dataset used is the T1DM Dataset, which considers the glucose level. The method is used to evaluate the Surveillance-error-grid (SEG) technique and the root-mean-square-error (RMSE) metric [14].

Jayanthi et al. [10] have concluded a predictive examination by using seven regression models. They include linear regression algorithm, polynomial regression, Lasso regression algorithm, logistic regression algorithm, stepwise regression, ridge regression and elastic net regression algorithm. In this paper, the idea of existing “predictive models” and “clinical predictive models” is given. In future, the truth of the existing system can be developed to a more by using many other predictive models and techniques[10].

Guolin et al. [4] has analyzed the concept of the LGBM Algorithm. They have used dual techniques : “Gradient Based one sided Sampling and the Theoritical Analysis” of the same. It includes exclusive Feature Bundling of large instances of data and produces results[4].

3. Basis of lgbm

There are many Machine Learning Algorithms that can be used to determine the accuracy of dataset involved. However, the LGBM Algorithm was considered to have the highest accuracy[17].

The LGBM Algorithm stands for Light Gradient Boosting Machine. The LGBM Algorithm mainly involves two concepts. They include GBDT (Gradient Boosting Decision Tree) or GOSS (Gradient based one-sided sampling). These algorithms are mainly used for prediction procedures of samples used during the study[17].

3.1 Gradient Boosting Decision Tree:

The GBDT is an algorithm that involves boosting technique. The concept of Boosting is a collective process that acts as a stronger classifier from a number of weaker classifiers involved. This can be obtained by building a model framework by working out the data included. It is then produced using a 2nd model that can be rectified from the mistakes included from the basic model framework created[31]. The efficiency, correctness and interoperability are its crucial factors that have to be measured. It is a growing tree data that is effective and accessible to all users[30] .

3.2 Gradient based one-sided sampling:

The other common algorithm that can be used for predicting is GOSS. Gradient-based One-Side Sampling (GOSS) can be used along with the LGBM Algorithm. When sampling is done, the GBDT whereas this can be overcome by the GOSS Algorithm. The down-sampling technique can be used for identifying the precision of the samples used[6].

3.1 Theoretical Analysis

GBDT uses the verdict trees concept along with a function and is given as follows: “from the input space X s to the gradient space G [6]. A training set with instances {x1, · · · , xn} are assumed, where each xi is a vector with dimension s in space X . In each restatement of gradient boosting, the negative gradients of the loss function with respect to the output of the model are denoted as {g1, · · · , gn}”[6]. “The decision tree model divides each node at the most revealing feature (which gives rise to the largest evidence gain). In GBDT, the data improvement is measured by the variance after segregating”, which can be explained as below[32].

“Y=Base_tree(X)-lr*Tree1(X)-lr*Tree2(X)-lr*Tree3(X)”

“Definition: Let O be the training dataset on a fixed node of the decision tree. The variance gain of dividing measure j at point d for this node is defined as

$$V_{j|O}(d) = \frac{1}{n_o}\left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \le d\}} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i\right)^2}{n^j_{r|O}(d)} \right)$$

where n$_O$ = $\sum I$[xi ∈ O], $n^j_{l|O}$(d) = $\sum I$[xi ∈ O : xij ≤ d] and $n^j_{r|O}$(d) = $\sum I$[xi ∈ O : xij > d]. This is ended by using the concept of GBDT”[22].

“The variance gain of dividing measure j at point d for this node is defined as

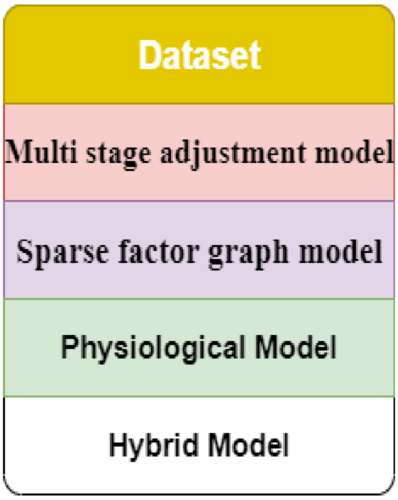
$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i + \frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{n^j_l(d)} + \frac{\left(\sum_{x_i\in A_r} g_i + \frac{1-a}{b}\sum_{x_i\in B_r} g_i\right)^2}{n^j_r(d)}\right)$$

where A_l = {xi ∈ A : xij ≤ d}, A_r = {xi ∈ A : xij > d},B_l = {xi ∈ B : xij ≤ d}, B_r = {xi ∈ B : xij > d}, and the coefficient $\frac{1-a}{b}$ is used to normalize the sum of the gradients over B back to the size of A^c. This is done by using the concept of GOSS”[6].

4. Prediction Models

The models that can be used for prediction of diabetes disease is

1. Multi stage adjustment model
2. Sparse factor graph model.
3. Physiological Model
4. Hybrid model



The multi stage adjustment model is used to identify the patients who are likely to be affected by the disease. The Sparse factor graph model are used to identify the underlying associations during the prediction procedure[2]. The physiological model is used to identify the blood glucose level in advance. The hybrid model is used to is used to identify if the patient is capable of being diagnosed by the disease within the next 5 years. It is also used to produce optimal feature subset[13].

3.3 Implementation of LightGBM

The LGBM Algorithm can be implemented well if the parameters considered are correct. There are over 100 parameters involved. The major parameters that are considered are as follows[21]:

Control - Parameters:

1. **Max Depth**: This is used to handle overfitting of the model. This determines the depth of the tree. In case there is a problem during execution, reducing the depth of the tree will reduce the error.
2. **Min data in leaf**: This determines the minimum number of records that the leaf is holding. This is also used for overfitting.
3. **Early Stopping around**: It is used to speed up analysis. This is used to remove the extra iterations during execution.

4. **Bagging Fraction**: The division of data for each iteration is done in this step. This is also used to overcome the problem of overfitting.
5. **Feature Fraction**: Used for random selection during iteration.
6. **Lambda**: This is used for regularizing the values.
7. **Min gain split**: It determines the minimum gain required to make in split in a tree.
8. **Max cat group**: It is used to identify the split points to group them into boundaries[20].

Core Parameters:

1. **Task**: It is used to determine the task that needs to be performed.
2. **Application:** It specifies the application of the model whether it belongs to classification or regression problem.
3. **Boosting:** It is used to determine the algorithm that is suitable.
4. **Learning Rate:** It examines the output received after execution depending on the effects caused on the tree.
5. **Number of leaves:** It states the maximum number of leaves a tree holds.
6. **Device:** This is default[28].

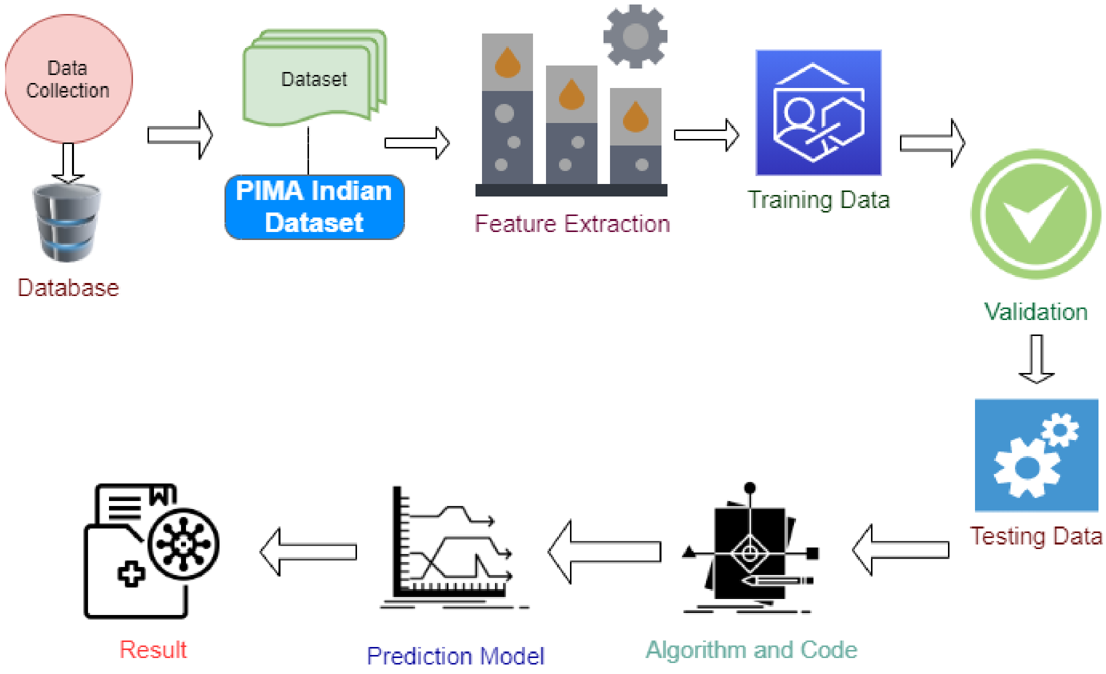
Metric Parameter:

1. **Metric**: It is used to identify values for mean squared error, mean absolute error, loss of binary classification and loss for multi classification procedures[26].

IO Parameter:

1. Max bin: It determines the maximum number of bins that will be used.
2. **Save Binary:** It is used for speeding the reading of the data. It is also used to save dataset to binary file.
3. **Ignore Column:** It is used to ignore specific columns.
4. **Categorical Feature:** It is used to denote the index of the features that are given into categories[26].

4. Architecture



The architecture of the system can be given analytically as follow: It starts with the process of data collection unit which is considered as the base cluster. It consists of all the details about the patients. Datasets containing historical medicinal data of a patient is significant. Due to this, the datasets can be obtained from databases available online. One such example of the database is Kaggle. A dataset that would be used is for the diabetic prediction is PIMA Indian Dataset for diabetes that can be taken from Kaggle. The other relevant datasets that are needed can be collected and consolidated to form the final dataset. The final dataset can be loaded into the IBM Watson cloud workspace after finalizing the dataset. It is run in a related machine erudition instance. This comes under the feature extraction and selection procedure. The machine learning instance verifies the dataset and extracts the fields that are unique in that particular dataset. This process is known as feature extraction.

Later, the structures that are extracted are now available for configuration of the user. The fields are now analyzed by selecting the fields using a prognostic algorithm and an output is obtained depending on the the relevant outcome.The arguments obtained are organized and this is known as feature selection. From the selected outcomes, grouping of data is done as training data and testing data. The training data produced a more accurate result when compared to the testing data. A validation procedure is included in order to filter the data further for easy testing.

The algorithm used to train and test the dataset is LGBM. It delivers maximum accuracy for the dataset tested when compared with the other algorithms involved. Following this procedure, the algorithm is selected and the dataset is divided into two parts as training data and testing data respectively. The calculation of data is done for the dataset taken. It didders based on the different algorithms used for predicting.

After the user splits the training and testing split, the algorithm will seslect the data to be used in the dataset and trains itself depending on the result obtained. The remaining data is also tested to predict. From the training and testing data, the final results are obtained to check into the prediction model which will give out the final obtained result.

After the analyzing the data using the predictive model, the results obtained can be observed and extracted in a variety of ways such as csv or xml or spreadsheet or word document as per the user's convenience. The results can be fine tuned by running the algorithm with any accessible developments and the accuracy of such enrichments can be pre-determined in the training and testing stage itself. Finally the accuracy of the dataset taken can be obtained.

5. Results and Discussion

In this research study conducted, the PIMA Indian Dataset is used the process of testing and validation of data. The algorithms used for comparison for the given dataset are Decision Tree, Extra Trees Classifier, Logistic Regression, XGB Classifier, Random Forest, Gradient Boosting Classifier and LGBM. The level of accuracy is checked as a percentage level for all the above mentioned algorithms and an observation that the LGBM Classifiers produce the best accuracy with a percentage of 95.20% is found out. The percentage of accuracy given by the other algorithms is given in the table below . Hence forth this is the final accuracy percentage for all the algorithms compared and studied for the taken dataset.

Dataset : PIMA Indian Dataset	**Logistic Regression**	**XGB Classifier**	**Gradient Boosting Classifier**	**Decision Tree**	**Extra Trees Classifier**	**Random Forest**	**LGBM**
	75.20%	83.30%	94.10%	94.40%	94.60%	94.80%	95.20%

6. Conclusion and Future Work

Therefore from the above discussions, we can observe that the LGBM Classifier Algorithm has produced the most accurate results by using the PIMA Indian Dataset. It can also be used to develop a data model in future, to predict and detect the Diabetes Mellitus Disease. In future, the LGBM algorithm can be further executed using Advanced LGBM Algorithm to castoff and bring out the necessary advancement in the field of medicine and for the Diabetes Mellitus Disease.

References

1. Ambigavathi et al., 2018 M. Ambigavathi, D. SridharanBig Data Analytics in HealthcareIEEE Tenth International Conference on Advanced Computing (ICoAC), (Dec 2018); pp-269-276.
2. Dash et al., 2019 S. Dash, S.K. Shakyawar, M. Sharma Big data in healthcare: management, analysis and future prospectsJournal of Big Data, Springer, 6, 54 (2019)
3. Goyal et al., 2021 J. Goyal, P. Khandnor & T.C.AseriA Comparative Analysis of Machine Learning classifiers for Dysphonia-based classification of Parkinson’s Disease.Int J Data Sci Anal 11, Springer,69–83 (2021)https://doi.org/10.1007/s41060-020-00234-0.
4. Guolin et al., 2017 Guolin Ke , Qi Meng , Thomas Finley , Taifeng Wang , Wei Chen , Weidong Ma1 , Qiwei Ye , Tie-Yan LiuLightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems (NIPS 2017), CA, USA.
5. Himansu et al., 2020 Himansu Das, Bighnaraj Naik, H.S. BeheraMedical disease analysis using neuro-fuzzy with feature extraction model for classificationInformatics in Medicine Unlocked, Volume 18, (2020), Pages 100299[Inf Med Unlocked 18 (2020) 1–12 page/100288]
6. Hosseini et al., 2020 M.M. Hosseini, M. Zargoush, F. AlemiLeveraging machine learning and big data for optimizing medication prescriptions in complex diseases: a case study in diabetes management. Journal of Big Data 7, Springer, 26 (2020) https://doi.org/10.1186/s40537-020-00302-z
7. https://thupilipraveenkumar.medium.com/what-is-lightgbm-how-to-implement-it-how-to-fine-tune - the-parameters-68b5ba7a76af.
8. Huzooree et al., 2017 G. Huzooree, K. K. Khedo and N. JoonasGlucose prediction data analytics for diabetic patients monitoring1st International Conference on Next Generation Computing Applications (NextComp), pp. 188-195, doi: 10.1109/NEXTCOMP.2017.8016197.
9. Huang et al., 2015 Z. Huang, W.Dong, P. BathOn mining latent treatment patterns from electronic medical records. Data Min Knowl Disc 29, Springer, 914–949 (2015) https://doi.org/10.1007/s10618-014-0381-y.
10. Jayanthi et al., 2017 N.Jayanthi, B. Vijaya Babu & N. Sambasiva RaoSurvey on clinical prediction models for diabetes prediction, Journal of Big Data volume 4, Article number: 26 (2017).
11. Kenneth David Strang 2020 Problems with research methods in medical device big data analyticsInt J Data Sci Anal 9, Springer, 229–240 (2020) https://doi.org/10.1007/s41060-019-00176-2
12. Kmar et al., 2019 S. Kumar and M. SinghBig data analytics for healthcare industry: impact, applications, and tools Big Data Mining and Analytics, vol. 2, no. 1, pp. 48-57, (March 2019)doi: 10.26599/BDMA.2018.9020031.
13. Kandhasamy et al., 2015 J.P. Kandhasamy, S. BalamuraliPerformance Analysis of Classifier Models to Predict Diabetes Mellitus Procedia Comput. Sci. (2015), 47, 45–51.
14. Martinsson et al., 2020 John Martinsson, Alexander Schliep, Björn Eliasson, Olof Mogren, Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks, Journal of Healthcare Informatics Research 4(2), (March 2020) DOI: 10.1007/s41666-019-00059-y.
15. Mamuda et al., 2017 M. Mamuda, S. SathasivamPredicting the survival of diabetes using neural networkProceedings of the AIP Conference Proceedings, 9–11 (May 2017); Vol 1870, pp. 40–46
16. Mercaldo et al., 2017 F. Mercaldo, V. Nardone, A. Santone,Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine LearniTechniques. Procedia Comput. Sci. (2017), 112, 2519–2528.
17. Magdelaine et al., 2015 Nicolas Magdelaine, Lucy Chaillous, Isabelle Guilhem, Jean-YvPoirier, Michel Krempf, Claude H Moog, Eric Le CarpentierA Long-Term Model of the Glucose–Insulin Dynamics of Type 1 DiabetesIEEE Transactions on Biomedical Engineering, vol. 62, no. 6, pp. 1546-1552, (June 2015)doi: 10.1109/TBME.2015.2394239.
18. Maecker et al., 2012 Maecker, Holden & Lindstrom, Tamsin & Robinson, William & Utz, Paul & Hale, Matthew & Boyd, Scott & Sheng, Deqiao & Fathman, CharlesNew tools for classification and monitoring of autoimmune diseases. Nature reviews. Rheumatology. 317-28. 10.1038/nrrheum.2012.66.
19. Nibareke et al., 2020 T. Nibareke, J. LaassiriUsing Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data 7, Springer, 78 (2020)https://doi.org/10.1186/s40537-020-00355-0
20. Negi et al., 2016 A. Negi, V.JaiswalA first attempt to develop a diabetes prediction method based on different global datasets

a. Proceedings of the (2016) Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), India, 22–24 (December 2016); pp. 237–241Olga Kolesnichenko 2019

21. Big Data Analytics of Inpatients Flow with Diabetes Mellitus type 1 : Revealing new awareness with Advanced Visualization of Medical Information System Data9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 191-196, doi: 10.1109/CONFLUENCE.2019.8776910
22. Olivera et al., 2017 A.R.Olivera, V. Roesler, C.Iochpe, M.I. Schmidt, A.Vigo, S.M.BarretoComparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes
23. ELSA-Brasil: accuracy study. Sao Paulo Med J. (2017 May-Jun);135(3):234-246.doi: 10.1590/1516-3180.2016.0309010217. PMID: 28746659.
24. Olaniyi et al., 2014 E.O. Olaniyi,K. AdnanOnset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng. Res. (2014),5, 754–759.
25. Rahul et al., 2020 Rahul Katarya, Sajal Jain Comparison of Different Machine Learning Models for diabetes detection IEEE International Conference on Advances and Developments in Electrical anElectronics Engineering (ICADEE) IEEE | DOI: 10.1109/ICADEE51157.2020.9368899
26. Sun et al., 2019 C. Sun, Q. Li, L. Cui, H. Li and Y. Shi Heterogeneous network-based chronic disease progression mining Journal of Big Data Mining and Analyticsvol. 2, no. 1, pp. 25-34, (March 2019), doi: 10.26599/BDMA.2018.9020009
27. Souad et al., 2019 Souad Larabi-Marie-Sainte, Linah Aburahmah , Rana Almohaini ,Tanzila Saba
    a. Current Techniques for Diabetes Prediction: Review and Case Study Applied Sciences, MDPI Journal,14 Published: (29 October 2019)
28. Sisodia et al. 2018 D.Sisodia, D.S SisodiaPrediction of Diabetes using Classification Algorithms.
    a. Procedia Comput. Sci. 2018, 132, 1578–1585.
29. Somnath et al., 2017 R. Somnath, M. Suvojit, B. Sanket, K. Riyanka, G. Priti, M. SayantanPrediction of Diabetes Type-II Using a Two-Class Neural NetworkProceedings of the (2017) International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 (March 2017); pp. 65–7.
30. Soltani et al., 2016 Z. Soltani, A. JafarianA New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II.Int. J. Adv. Comput. Sci. Appl. (2016), 7, 89–94.
31. Thulasi Bikku 2020 Multi-layered deep learning perceptron approach for health risk predictionJournal of Big Data 7, Springer, 50 (2020)https://doi.org/10.1186/s40537-020-00316-7
32. Tafa et al., 2015 Z. Tafa, N. Pervetica, B. KarahodaAn intelligent system for diabetes predictionProceedings of the (2015) 4th Mediterranean Conference on Embedded Computing (MECO),Montenegro, 14–18 (June 2015); pp. 378–382.
33. Ukil et al., 2016 A. Ukil, S. Bandyoapdhyay, C. Puri and A. PalIoT Healthcare Analytics: The Importance of Anomaly Detection30th International Conference oAdvanced Information Networking and Applications (AINA), pp. 994-997, doi: 10.1109/AINA.2016.158.
34. Wang et al., 2015 F. Wang, G. Stiglic, Z. ObradovicGuest editorial: Special issue on data mining for medicine and healthcare. Data Min Knowl Disc 29, Springer, 867–870 (2015) https://doi.org/10.1007/s10618-015-0414-1.
35. Yuvaraj et al., 2017 N. Yuvaraj, K.R. SriPreethaaDiabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust. Comput. (2017), 22, 1–9.