Visualizing the Clinical Data of Diabetes using Data Science and Machine Learning Algorithms

Vennapusa Vishnu Priya, Abdul Gaffar.H

***Abstract**: In recent decades Machine learning and Data Science are providing best ways to analyze and solve various problems. In fact, those Machine Learning algorithms gives the best and optimized solutions. These methods are playing key role in providing efficient solutions for the health care problems like predicting the diseases in early stage, and even some automated systems run by Machine Learning Algorithms are prescribing medicines based on the patient's symptoms. Diabetes is one among the chronic diseases from past years, which leads to the damage of patients eyes, nerves, heart and kidneys etc., In this project we are going to create a pipeline in which the data collected from the source is undergone through some preprocessing techniques and the Machine Learning Algorithms like SVM, KNN, Gradient Boasting, logistic regression and Random Forest are used to classify whether the patient is diabetic or not and the accuracy of these algorithms was measured by using some Evaluation methods like Train/Test Split. Finally, these data will be visualized by using Visualization Tools.*

***Index Terms**: SVM, KNN, Gradient Boasting, Logistic regression, Random Forest.*

I. INTRODUCTION

In current years medical knowledge is increasing immensely in varied dimensions. Because the knowledge is unceasingly increasing during this field, managing this knowledge are going to be a giant task. D.M is one of the long-standing diseases. There are 3 sorts of diabetes: sort 1-typically happens in youngsters, sort 2- of times happens in adults, and kind 3- happens in pregnant ladies.

An estimation of 387 individuals everywhere the globe area unit tormented by polygenic disease, among them concerning 90 area unit of kind of 2 pair of polygenic disease.

Approximately 2-5 million deaths can occur once a year due to polygenic disease.

By 2035 the amount of individuals touching polygenic disease could get up to 592 million.

When involves information science, it's a specialized field th at mixes multiple areas like statistics, arithmetic intelligent information capturetechniques, information cleans ing, mining and programming to arrange and align huge information for intelligent analysis to extract insights and data, although this is going to sound straight forward information science is sort of a difficult space thanks to the complexities concerned in combining and applying totally different ways, algorithms and sophisticated programming techniques to perform intelligent analysis in massive volumes of knowledge Hence, the sector science has evolved from huge data, or huge information and information science area unit indivisible. However, there are a unit several variations between huge information and information science. The main objective of this project is to predict whether the patient is diabetic or not. For the prediction we tend to use some supervised Machine Learning algorithms like SVM, KNN, Gradient Boasting, logistic regression and Random Forest not and the accuracy of these algorithms was measured by using some Evaluation methods like Train/Test Split. And conjointly to envision the complete knowledge by employing a visualization tool.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author

Vennapusa Vishnu Priya*, SCOPE, Vellore Institute of Technology, Vellore, India.
Prof.Abdul Gaffar H, SCOPE, Vellore Institute of Technology, Vellore, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

II. LITERATURE SURVEY

A. Diabetes Prediction Using Machine Learning Techniques

In this paper authors aim was to improve the accuracy of the classification and to classify whether the data is diabetic positive or diabetic negative. Here they had selected many numbers of samples for the classification purpose, but this doesn't provide the better accuracy. In most of the cases, the performance rate is high but not the classification speed and accuracy. The main aim of this paper is to achieve a model of high accuracy. They did survey and analyzed various techniques for classification and observed that, some techniques like SVM, Artificial Neural Networks and Logistic Regression are suitable for the prediction of diabetes.

B. Prediction of Diabetes using Classification Algorithms

In this paper, the authors made their efforts to design a model to predict the diabetes. In this work they studied and evaluated some ML algorithms on various measures. In the system designed by them, they used three classification algorithms among which the Naive Bayes algorithm have given 76.3% of accuracy. Here the authors concluded that the ML classification algorithms can be used for the prediction of various diseases, their work can be extended and can be improved for the analysis of diabetes by using some other ML algorithms.

C. *Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics*

In recent years Big Data and Cloud is playing a major role in solving the problems of healthcare. As the healthcare data is continuously increasing and it is necessary to manage this data efficiently. One of the chronic healthcare issue is diabetes, even some times this leads to death. In this paper the author is trying to identify the best predicting algorithm of machine learning based on some matrices like Kappa, Sensitivity, accuracy and specificity. A study was done on the dataset of diabetes with various ML algorithms. Based on the results RF algorithms produces the better predictions.

D. *Machine learning techniques for classification of diabetes and cardiovascular diseases*

In this paper the creator gives a diagram of arrangement of diabetes and cardiovascular maladies (CVD) by utilizing AI calculations like utilizing Artificial Neural Networks (ANNs) and Bayesian Networks (BNs). Here they did relative investigation on chose papers which are distributed in the period from 2008 to 2017. Among the chose papers ordinarily utilized sort of ANN is multilayer feed forward neural system with Lederberg-Marquardt learning calculation. The most noteworthy arrangement exactness esteem is 99.51% which was accomplished by the Naive Bayesian system. Amid the arrangement, the mean precision of watched systems has given great outcomes by utilizing ANN, this demonstrates the likelihood to get progressively exact outcomes when it is connected to ANN.

III. Machine Learning

AI is a use of man-made brainpower (AI) that gives capacity to the frameworks to take in consequently and improve from the encounters without being unequivocally modified. AI centers around advancement of PC programs which can get to information and use it to learn without anyone else's input.

The way toward learning starts with perceptions or information, for example, models, direct understanding, or guidance, as to search for the examples in information and to settle on better choices later on dependent on the precedents that we give. The principle point is to influence the PCs to adapt consequently independent from anyone else without human help or intercession and modify activities in like manner. Tremendous measure of information can be empowered by investigation of the Machine learning. While typically it conveys progressively precise outcomes quicker, so as to recognize the better chances or perilous dangers, it might likewise require extra assets and time to prepare it legitimately. The mix of AI with the AI and psychological advances can give adequacy in handling of expansive volumes of information.

IV. ALGORITHMS

A. *Support Vector Machine*

Support vector machine is another straightforward formula that each machine learning knowledgeable ought to have in his/her arsenal. Support vector machine is extremely most popular by several because it produces vital accuracy with less computation power. SupportVectorMachine, abbreviated as SVM will be used for each regression and classification tasks. But, its wide utilized in classification objectives. The objective of the support vector machine formula is to search out a hyper plane in Associate in Nursing N-dimensional space (N — the range of features) that clearly classifies the info points. To separate the 2 categories of knowledge points, there are several attainable hyperplanes that would be chosen. Our objective is to search out a plane that has the utmost margin, i.e the utmost distance between information points of each categories. Maximizing the margin distance providessome reinforcement so future information points will be classified with additional confidence.Hyper planes are call boundaries that facilitate classify the info points. Information points falling on either facet of the hyper plane will be attributed to totally different categories. If the amount of input options is three, then the hyper plane becomes a two-dimensional plane. It becomes tough to imagine once the amount of options exceeds three. Support vectors are data points that are nearer to the hyper plane and influence the position and orientation of the hyper plane. Practice these support vectors, we&#39;ve got a bent to maximize the margin of the classifier. Deleting the support vectors will modification the position of the hyper plane. These are the points that facilitate North yank country build our SVM.

B. *Logistic Regression*

Logistic regression is that the most known machine learning rule once rectilinear regression. in a very ton of the way, rectilinear regression and supply regression area unit similar. The imp distinction lies in rectilinear regression algorithms area unit accustomed predict/forecast values however supply regression is employed for classification tasks. If you're shaky on the ideas of rectilinear regression. There are a unit several classification tasks done habitually by folks. for instance, classifying whether or not associate degree email could be a spam or not, classifying whether or not a growth is malignant or benign, classifying whether or not a web site is deceitful or not, etc. These area unit typical examples wherever machine learning algorithms will build our lives a great deal easier. A very straightforward, rudimental and helpful rule for classification is that the supply regression rule.

C. *K-Nearest Neighbor (K-NN) Classifier*

A k-nearest-neighbor could be a knowledge classification rule that tries to work out what blood group information is in by gazing the info points around it. An algorithmic rule, staring at one purpose on a grid, attempting to see if a degree is in type A or B, appearance at the states of the points that square measure closes to it. The range is randomly determined; however the purpose is to require a sample of the information. If the bulk of the points area unit in type A, then it's probably that the information purpose in question are A instead of B, and vice versa.

The k-nearest-neighbor is associate degree example of a & quot; lazy learner&quot; algorithmic rule as a result of it doesn't generate a model of the information set beforehand.

1. *What is K?*

For each check information, we might be observing the K nearest knowledge points and take the foremost often occurring categories and assign that category to the check knowledge.

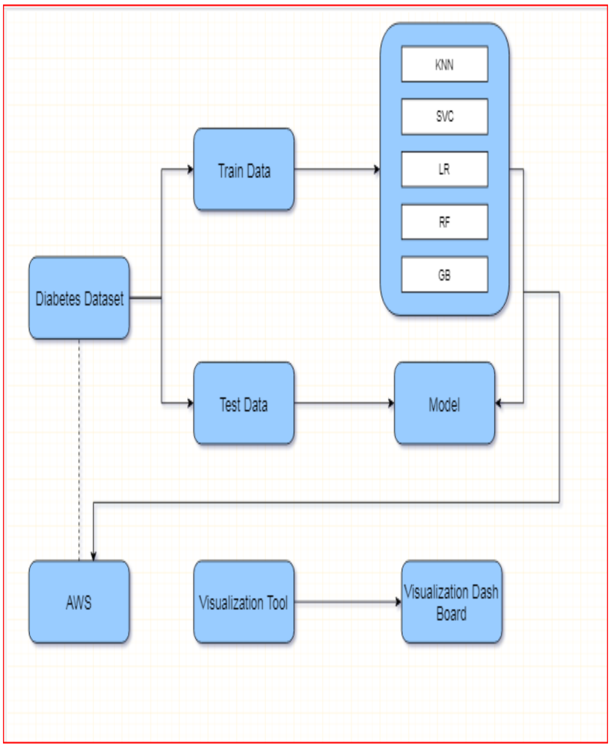
Therefore, K represents the quantity of coaching purpose lying in proximity to the check information point that we have a tendency to area unit attending to use to search out the category.

V. Data Visualization

Data visualization normally describes a kind of effort that helps the people to understand the importance of data by showing it in a visual context. The features that cannot be identified in the text-based data, can be recognized and exposed easily with data visualization software.

Today's data visualization tools go beyond the graphs and standard charts which are used in Microsoft Excel spreadsheets, displaying the data in more refined ways such as geographic maps, spark lines, info graphics, gauges and dials, heat maps, and detailed bar, fever and pie charts. The plots also include various interactive capabilities that enables the users to manipulate or drill into the data for analysis and querying. When predefined conditions occur, or data has been updated can be send as alerts to the users by the designed indicators.

VI. PROPOSED SYSTEM



The data was separated in to train data and test data. Some supervised machine learning algorithms like SVM, KNN, Gradient Boasting, logistic regression and Random Forest are used for the prediction. Like all other machine learning models, the model was first trained by the train data. Here the accuracy of these algorithms was measured by using some Evaluation methods like Train/Test Split. The predicted output and the data was pulled from the EC2 instance, Then the data was visualized by using the Visualization tools, Here we are using plotly chart studio for the visualization purpose and finally the visualization dash board was created. The dashboard will contain various types of plots which are plotted from the data.

VII. CONCLUSION

In this project we build a visualization pipe line. Based on the pipeline we are going to build a classification model to classify weather diabetic or not. Here the accuracy is measured by Train/Test Split with Scikit Learn. Then the data was visualized by using plotly chart studio. A dashboard was plotted with various plots.

VIII. FEATURE SCOPE

Here in this paper we are predicting whether the patient is diabetic or not and visualizing the data. Even this prediction model can be built for some other chronic diseases like cancer, heart stroke etc., and we can visualize the entire hospital data in a dash board in which we can select a disease in the dropdown to display. In this model we used some Supervised Machine Learning algorithms and plotly chart studio for visualization instead of these we can use some other advanced algorithms and visualization techniques for better out puts. These outputs can be sent as notifications by mails for the concerned persons. This paper can be developed as a good real time project in feature.

IX. RESULT Analysis

The clinical data of diabetes is collected, then some ML algorithms are used to classify the types of diabetes and Visualized. Here for visualization we are using plotly chart studio. Other Recommendations.

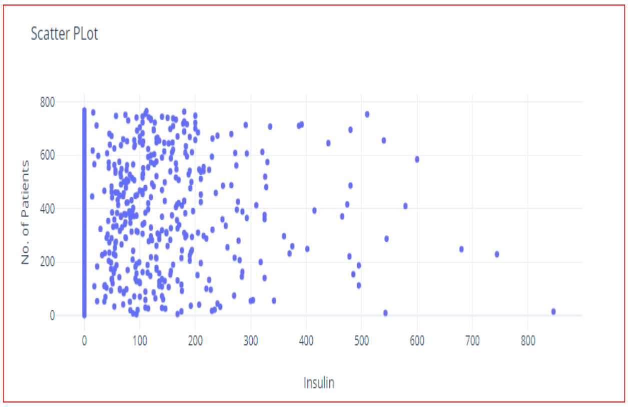


Figure 1 scatter plot

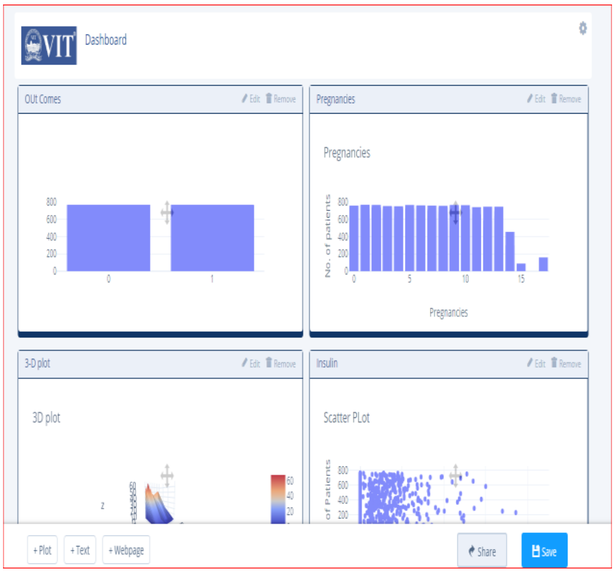


Figure 2 Dashboard

REFERENCES

1. Predictive Analytics in Health Care Using Machine Learning Tools and Techniques by B.NIthya and Dr.V.Ilango presented on International Conference on Intelligent Computing and Control Systems ICICCS 2017
2. Machine learning techniques for classification of diabetes and cardiovascular diseasesby Berina, LejlaeGurbeta, Almir presented in 20 17 6th MEDITERRANEAN CONFERENCE ON EMBEDDED COMPUTING „/" (MECO), 11-15 JUNE 2017, BAR, MONTENEGRO.
3. Geo-Identification of Web Users through Logs using ELK Stack by Tarun Prakash, Ms. Misha Kakkar and Kritika Patel published on 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence).
4. Artyom Churilin, "Choosing an open-source log management system for small business," Master's Thesis, Faculty of Information Technology, Tallin University of Technology, Tallinn, Estonia.
5. N. Sandhya, K.R. Charanjeet, A review on Machine Learning Techniques, International Journal on Recent and Innovation Trends in Computing and Communication, 2016, ISSN: 2321-8169, 395 – 399
6. Christopher (2015, April 15).Visualizing data with Elasticsearch, LogstashandKibana[Online].Available:http://blog.webkid.io/visualize datasets-with-elk/
7. Anders Aarvik (2014, April 04).A bit on ElasticSearch + Logstash +Kibana(TheELKstack)[Online].Available:http://aarvik.dk/a-bit-onel asticsearch-logstash-kibana-the-elk-stack/
8. GeoLiteLegacyDownloadableDatabase[Online].Avaialble:http://dev. maxmind.com/geoip/leg acy/geolite/
9. A. Ghaheri, S. Shoar, M. Naderan and S.S. Hoseini, The applications of genetic algorithms in medicine. Oman medical journal, 2015, 30(6), 406

AUTHORS PROFILE

Vennapusa Vishnu Priya, MTech Computer Science and Engineering, SCOPE, Vellore Institute Of Technology, Vellore, INDIA. Research on Data Science and Visualization.

Prof Abdul Gaffar.H, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, INDIA.