PREDICTIVE ANALYSIS ON DIABETES USING SUPERVISED MACHINE LEARNING ALGORITHMS

Roopasri K[a] and Bharanidharan A[b]

***Abstract*— Diabetes mellitus (DM) results due to insulin deficiencies, which in turn lead to chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism. The existing testing system requires a multiple lab assessments and it is a tedious process. Machine learning and computational intelligence techniques play a vital role in transforming healthcare which provides objective decision support tools to assist medical professionals in diagnosis and prognosis the patient conditions. Smart machine learning algorithms are used now-a-days in different industries to replace the costly, repetitive and time consuming tasks. They also capture unforeseen patterns within the complex data set at must faster rate which would have not been seen by human eye and brain. In the present work, pima Indians diabetes data set is taken and employed in the three different machine learning algorithms namely Support Vector Machine (SVM) algorithm, K-Nearest Neighbour (KNN) and Naïve bayes. KNN algorithm is found to have the best accuracy and most suited algorithm. The data set are further analyzed using graph maker - plotly and is observed that around 20-30 years of age groups found to have high glucose level and preventive measures to be taken in the young age itself.**

***Index Terms*— Diabetes, KNN, Machine learning, Naïve bayes, Supervised learning, SVM**

I. INTRODUCTION

In recent years, there is a tremendous interest in the usage of machine learning in various fields such as virtual personal assistants, predictions while commuting, video surveillance, social media services, email spam and malware filtering, medical diagnostics etc., Machine learning and computational intelligence techniques play a vital role in transforming healthcare that provides objective decision support tools to assist medical professionals in diagnosis and prognosis the patient conditions[1].

Diabetes mellitus (DM) is a group of metabolic disorders resulting from a defect in insulin secretion, insulin action, or both [2]. Insulin deficiencies in turn lead to chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism. It is one of the chronic (life long) disease and is caused due to increase in blood sugar. As the disease progresses tissue or vascular damage ensues leading to severe diabetic complications such as retinopathy, neuropathy, nephropathy, cardiovascular complications and ulceration. Thus, diabetes covers a wide range of heterogeneous diseases. It is the most common endocrine disorder and by the year 2025, it is estimated that more than 300 million people worldwide will have DM [3-5].

Machine Learning technique is used to make prediction by training dataset based on different algorithms. Smart machine learning algorithms are used now-a-days in different industries to replace the costly, repetitive and time consuming tasks [6]. They also capture unforeseen patterns within the complex data set at must faster rate which would have not been seen by human eye and brain. These algorithms provides all sorts of forecast such as the possibility of developing a certain conditions after a period, need for re-hospitalization after some time and even patient's response to new drug.

Self learning by machines has created more interest in medical domain because machines are fed with data, analyze it and suspect the chance of being prone to disease [7]. These results after being evaluated helps doctor to recommend patients the chance of being diseased and to take preventive measures in order to avoid further complications and detect it in earlier stage.

In the present work, pima Indians diabetes data set is taken and employed in the three different machine learning algorithms namely Support Vector Machine (SVM) algorithm, K-Nearest Neighbour (KNN) and Naïve bayes and the results and predictions are discussed.

II. LITERATURE REVIEW

Ioannis Kavakiotis et al.,[8] have reported about the machine learning and data mining methods in diabetes research. They have conducted a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to Prediction and Diagnosis, Diabetic Complications, Genetic Background and Environment, and Health Care and Management with the first category appearing to be the most popular. A wide range of machine learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector

[a] III B.E, Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India.
Email : roopasrivk@gmail.com
[b]Assistant Professor, Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India.
Email : bharanidharan@srec.ac.in

machines (SVM) arise as the most successful and widely used algorithm.

Saravana Kumar et al.,[9] have reported the predictive methodology for diabetic data analysis in Big data. They used the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, their system provided an efficient way to cure and care the patients with better outcomes like affordability and availability.

Ravi Sanakal et al., [10] have reported the implementation of FCM and SVM and testing it on a set of PIDD which gives good classification. It also stated better machine learning algorithm should be employed along with them.

Humar Kahramanli et al.,[11] have worked on Artificial neural network combined with fuzzy logic which was used to detect diabetes. It allows better result as fuzzy accounts for uncertainties also. But extracting rules from existing methods was not very efficient as it takes time.

B.M. Patil et al., [12] have proposed Hybrid prediction model for Type-2 diabetic patients in which simple K-means clustering algorithm was used. C4.5 algorithm was used to build the final classifier. It was also studied that Hybrid approach gives better result as compared to single classifiers.

Mani Butwall et al.,[13] reported that Data mining approach to envisage diabetes behaviour is based on Random Forest Classifier was it was good approach to handle large data set. But single classifier approach was not very effective as compared to hybrid.

Nawaz Mohamudally et al.,[14] have considered Neural Network, Kmeans, Visualization was used to detect diabetes. It was good approach as hybrid method was used. But the major drawback was prediction, classification, visualization requires tremendous effort.

Veena Vijayan V et al.,[15] have proposed that Decision support system uses AdaBoost algorithm with Decision Stump as base classifier for classification. Support Vector Machine, NaiveBayes and Decision Tree are also implemented as base classifiers. Eventhough Adaboost gives an edge to yield combined and better results; Accuracy of classifiers needs to be improved with nn classifiers and other approaches.

Kiarash ZahirniaMehdi et al.,[16] have presented the comparison of different cost-sensitive learning methods for diagnosis of type 2 diabetes. As Cost sensitive approach was effective for utilizing resources, the drawback was the assumptions used in data sets, matrices to bring out the results.

III. SUPERVISED MACHINE LEARNING ALGORITHMS

SVM

Support Vector Machine (SVM) algorithm is a method which performs classification tasks. In multidimensional space, hyper planes also said to be decision boundaries are constructed that are used to separate different class labels. It handles both categorical and continuous variables [17].

Linear SVM $x_i \cdot x_j$ (1)

Non-Linear SVM $\phi(xi) \cdot \phi(xj)$ (2)

Kernel function $k(xi \cdot xj)$ (3)

KNN

K-Nearest Neighbors is a simple algorithm, which uses the entire dataset as the training set, rather than splitting the dataset into a training set and test set. It classifies a new instance based on distance function. These formulas are used only for continuous variables [18].

Euclidean $\sqrt{\sum_{i=1}^{k}(x_{i-}y_i)^2}$ (4)

Manhattan $\sum_{i=1}^{k}\left|x_i - y_i\right|$ (5)

Minkowski $\left(\sum_{i=1}^{k}\left(\left|x_i - y_i\right|\right)^q\right)^{1/q}$ (6)

NAÏVE BAYES

Naïve bayes theorem is used to calculate the probability that an event will occur, given that another event has already occurred [18].

$$P(h|d) = (P(d|h) * P(h)) / P(d) \quad (7)$$

where

- P(h|d) = Posterior probability. The probability of hypothesis h being true, given the data d, where P(h|d)= P(d1| h)* P(d2| h)*....*P(dn| h)* P(d)
- P(d|h) = Likelihood. The probability of data d given that the hypothesis h was true.
- P(h) = Class prior probability. The probability of hypothesis h being true (irrespective of the data)
- P(d) = Predictor prior probability. Probability of the data (irrespective of the hypothesis)

IV. PROPOSED SYSTEM

The proposed system mainly focuses to find the suitable machine learning algorithm with best accuracy for the diabetes dataset taken. By employing this technique, three classifier models were built using Anaconda's Jupyter Notebook, which is a powerful tool for Data Analysis. Jupyter Notebook is built of *IPython*, which interact and run Python code in the terminal using the *REPL model* (Read-Eval-Print-Loop). Further the dataset was analyzed using Graph maker Plotly online to obtain scatter plot graph for the dataset taken to confirm the same.

V. RESULTS AND DISCUSSION

The on-line pima Indians – diabetes data set is taken from kaggle for the predictive analysis of prevalence of diabetes for the Indians. Three different machine learning algorithms namely SVM algorithm, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) are employed for this data. The results are compared for the algorithms as shown in Figure 1.

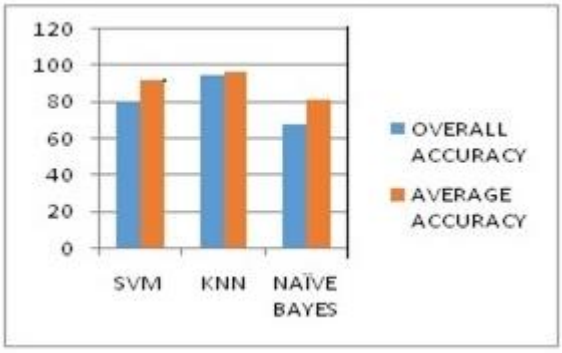

Figure 1.Comparative performance of ML algorithm

It is observed from the results that the overall accuracy is found to 79 for SVM algorithm, 96 for KNN algorithm and 67 for Naïve Bayes algorithm. It is further observed that KNN algorithm is found to be the most suited algorithm for the given data set [19]. The datas can be further analyzed using graph maker plotly and their results are shown in Figures 2-5.

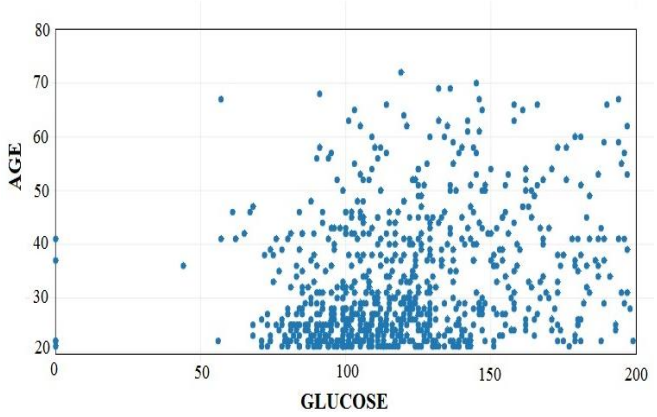

Figure 2. Scatter Plot of Glucose Vs Age

The graph is plotted using graph maker software to analyze the glucose levels for the different age group to predict the presence of diabetics. The graph is plotted taking glucose level in the X-axis and Age in the Y-axis, by taking available online dataset as shown in Figure 2. It is observed from the graph that, around the age group of 20-30 years, the glucose level is found to be high. Hence preventive measures to be taken in the young age itself.

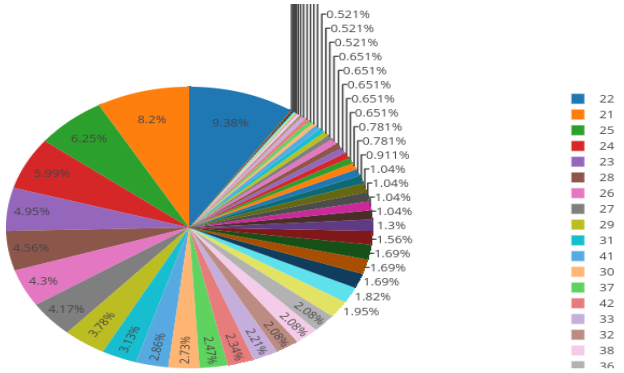

Figure 3. Pie chart of Age

Further pie graph also ensures that diabetics are found to be high in the age group of 22 years as shown in Figure 3. The accuracy result could be obtained from the pie chart, which can be used to support the prediction had from the machine learning algorithms.

Further, Body mass index is a strong and independent risk factor for being diagnosed with type diabetes mellitus. Type 2 diabetes risk may be incrementally higher in those with a higher body mass index. Understanding the risk factors helps to shorten the time to diagnosis and treatment. Hence the graph is plotted taking BMI along the X-axis and Age along the Y-axis as shown in Figure 4.

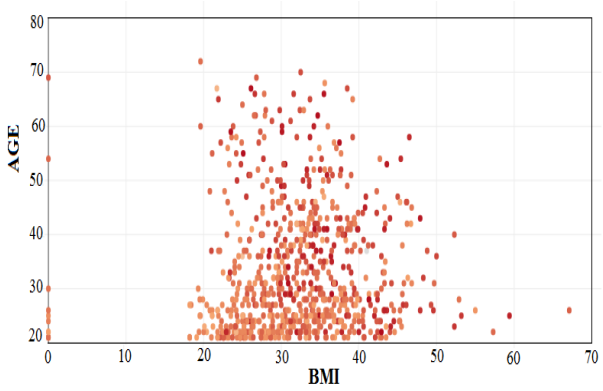

Figure 4. Scatter Plot of BMI Vs Age

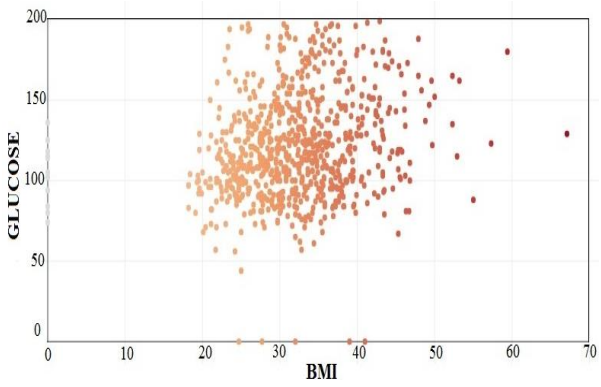

Figure 5. Scatter Plot of BMI Vs Glucose

The scatter plot graph also predicts that BMI is found to be high for the age group of 25-31 years for the data taken from

online dataset. The relation between glucose and BMI can be further confirmed by plotting BMI along the X-axis and Glucose along the Y-axis as shown in Figure 5.

VI. CONCLUSION

Diabetes is one of the chronic diseases and is caused due to increase in blood sugar. It is estimated that more than 200 million people worldwide will have Diabetes Mellitus and 300 million will subsequently have the disease by 2025. Hence in this project, machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Naive-Bayes have been employed to analyze and predict from large set of patient records taken from online data set. It is found that KNN algorithm has the best accuracy compared to Naïve-Bayes and support vector machine algorithms. The data sets are further analyzed using graph maker - plotly and are observed that around 20-30 years of age groups found to have high glucose level and preventive measures to be taken in the young age itself. In near future, it can be implemented with other algorithms and can be applied for real data sets.

REFERENCES

1. http://www.cibcb2017.org/Machine_Learning_in_Medical_Diagnosis_and_Prognosis.pdf
2. Ann M. Aring, David E. Jones and M. James "Falko Evaluation and Prevention of Diabetic Neuropathy", Am Fam Physician, vol. 71, pp. 2123-2128, 2005.
3. A. Amos , D. McCarty and P. Zimmet, "The rising global burden of diabetes and its complications, estimates and projections to the year 2010",Diabetic Med., vol. 14, pp.S1- S85, 1997.
4. H.King, R.Aubert and W.Herman, "Global burden of diabetes, 1995-2025, Prevalence, numerical estimates and projections", Diabetes Care, vol. 21, pp.1414-1434, 1998.
5. P.Zimmet, "Globalization, coca-colonization and the chronic disease epidemic: can the Doomsday scenario be averted?", J Med, vol.247, pp.301-310, 2000.
6. www.romexsoft.com/blog/improve-medical-diagnosis-using-machine-learning/
7. www.techemergence.com/machine-learning-in-pharma-medicine/
8. Ioannis Kavakiotis, OlgaTsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology, vol.15, pp.104-116, 2017.
9. N.M.Saravanakumar, T.Eswari and P. Sampath and S.Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data", Science Direct ,vol. 50, pp.203, 2015.
10. Ravi Sanakal and T. Jayakumari, "Prognosis of Diabetes Using Data mining Approach -Fuzzy C Means Clustering and Support Vector Machine", International Journal of Computer Trends and Technology (IJCTT), vol.11, pp.94-98, May 2014.
11. Kahramanli, Humar and NovruzAllahverdi, "The performance comparison of discrete wavelet neural network and discrete wavelet adaptive network based fuzzy inference system for digital modulation recognition", Expert Systems with Applications, vol.35, pp.90-101, 2008.
12. Patil, M.Bankat, Ramesh Chandra Joshi and DurgaToshniwal, "Hybrid prediction model for Type-2 diabetic patients", Expert Systems with Applications, vol.37, pp.8102-8108, 2010.
13. Butwall Mani and Shraddha Kumar, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, vol.120, pp.36-39, 2015.
14. Mohamudally Nawaz, and Dost Muhammad Khan, "Application of a unified medical data miner (umdm) for prediction, classification, interpretation and visualization on medical datasets: The diabetes dataset case", published in Industrial Conference on Data Mining - Advances in Data Mining. Applications and Theoretical Aspects, 2011, pp-78-79.
15. V.VeenaVijayan and C.Anjali, "Prediction and diagnosis of diabetes mellitus—A machine learning approach", Intelligent Computational Systems (RAICS), 2015.
16. Kiarash, Zahirnia Mehdi, TeimouriRohallah, Rahmaniand Amin, Salaq, "Diagnosis of type 2 diabetes using cost-sensitive learning" published in Computer and Knowledge Engineering (ICCKE), 2015 5th International Conference on. IEEE, 2015.
17. www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
18. https://gerardnico.com/data_mining/knn/naive_bayes.
19. Priya B.Patel, ParthP.Shah and HimanshuD.Patel, "Analyze Data Mining Algorithms for Prediction of Diabetes", IJEDR, vol. 5, pp.466-473, July 2017.