Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction

OMOSHOLA OGUNDUBOYE
Deparmtent of Data Science and Artificial Intelligence, Bournemouth University

Abstract- ***Diabetes is regarded as one of the most chronic metabolic diseases if left unchecked. Diabetes is a worldwide chronic health issue. Today, approximately 400 million people are living with diabetes. A large percentage of people who are living with diabetes are unaware of their condition until it becomes chronic. Diabetes, also known as Diabetes Mellitus, is an increasingly prevalent chronic disease which affects the body's ability to metabolize glucose. With the growing rate of diabetes cases, it has become important to take a deeper look into solutions and ways to better handle the situation. This paper presents a predictive approach to diabetes, through diabetes prediction using machine learning, a process that will allow for better treatment and preventive healthcare. Machine learning in diabetes prediction is important because there is a vast pool of available data on diabetes both through research and years of clinical studies. This data can be processed and fed into machine learning models to highlight meaningful relationships and patterns within patients' data. However, this has been hampered by the difficult task of choosing the best machine learning algorithm. A challenge that can be solved by carrying out a comparative study using different evaluation metrics to ascertain which algorithm produces the most optimal results. This paper represents the result and analysis regarding detecting a person's diabetic state from various machine learning models based on key attributes such as age, gender, glucose level and insulin level. The model proposed was achieved by collating diabetes data from Kaggle and prepossessed to remove abnormalities and irrelevant attributes after which it was divided into test and training data. The machine learning algorithms chosen for this study were SVM, logistic regression, decision tree, random forest classifier and* **K-Neighbors** *classifier. The best performing model was random forest with an accuracy of 95%. This paper contributes to the diagnosis and prediction of diabetes through the application of machine learning in predicting patients who are likely to live with diabetes.***

Indexed Terms- ***Diabetes, Decision Tree, Dataset, Attributes, Machine Learning, SVM, K-Neighbors, Random Forest Algorithms.***

I. INTRODUCTION

Machine learning in healthcare has witnessed exponential growth over the last decade [1], this growth has without doubt proven that machine learning has the potential to transform the way healthcare is delivered. With the growth of machine learning's application to healthcare, the number of research initiatives and studies towards growing the body of work also increases, but there is still a limited knowledge base on how AI and machine learning can and should be applied to the various fields of healthcare [2] and how this would affect heath care.

Machine learning can be regarded as one of the most disruptive technologies of recent times [3], this is due to its vast applications, in almost all aspects of our daily lives as well as applications in other major fields. Machine learning is an important concept because it involves the development of algorithms which can learn from the available data and make decisions and predictions on their own. The key parameter here is the availability of data, for an AI-enabled machine to progress, evolve or learn, there must be data available, without data learning is not possible. With data available, machines can self-learn, make corrections and learn from past errors and experiences. In other words, machine learning is the extraction of knowledge and

information from data for making decisions. The primary goal of machine learning is to identify patterns in raw data, then perform useful predictions using the patterns that it has already learnt from. The key concept here is the use of machine learning algorithms as opposed to traditional algorithms. Machine learning algorithms learn and identify patterns in available data and uses the knowledge to predict patterns in the new data in a similar way. This process does not require explicit programming on how to execute the task. In contrast to machine learning algorithms, traditional algorithms involve a manual process of creating a program and explicitly conditioning the rules and execution process, it also involves feeding the program an input which produces a known or desired output. Figure 1 shows how traditional programming differs from machine learning.

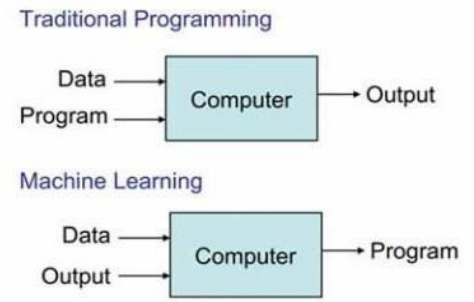

Fig 1: Diagram of Traditional and Machine Learning [4]

Diabetes mellitus is a chronic disease characterized by high blood sugar which often leads to metabolic complications. With the growing rate of diabetes cases worldwide, there is a need for a viable solution [5].

In recent times the terminology “big data” is becoming exceedingly popular all over the globe [6]. This is due to the numerous applications of big data in day-to-day human activities. Following that, scientists have been working on improving the healthcare, medical research, and the care provided to patients by analyzing big datasets related to their health. The data is being sourced from medical records, hospitals and patients’ history, web searches, even interactions on social media, a very popular source of big data is Twitter. The healthcare sector stands to benefit more from big data, with the growing number of diabetes cases, there is need for a level of automation in prognosis and diagnosis [7].

Healthcare institutions have a wide pool of data available to them through years of medical records [8], clinical trials and journals. Big data is just an aspect of what is needed to aid in the development of health prediction models, because the data obtained must be interpreted correctly to predict future cases of diabetes cases with accuracy. To solve this problem, there is a need for data mining for the analysis of huge quantities of raw data and extraction of useful information from it.

Data Mining is a field that is based on various fields including artificial intelligence [9], high-performance computing, visualization, statistics, pattern recognition, neural networks and machine learning. The applications of data mining in healthcare are becoming more mainstream and popular. Data mining is very vital to the healthcare sector allowing for improved care, personalized health care [10], better diagnosis, aiding better healthcare services, affordable and personalized healthcare [11]. Data mining also shows great opportunities for hidden pattern explorations from large data. These patterns can be used by doctors to establish diagnoses, prognoses and treatment for patients in healthcare institutions [12]. The application of data mining in this study is to aid the prediction of the possibilities of diabetes in patients. The study uses diabetes dataset sourced from Kaggle, an open-source data platform to build a model that can be used for diabetes prediction.

This study presents a comparative analysis of machine learning techniques for diabetes prediction. The rest of this study is organized as follows; Section 2 introduces other researcher’s contributions to the body of work. In section 3, the methodology of the study is defined. The results of the evaluation study are illustrated in section 4. Finally, the concluding note of the

study is presented in section 5.

II. LITERATURE REVIEW

This section introduces other researchers' work within the field of diabetes prediction. The analysis of other researchers' work provides an in-depth knowledge of possible gaps and findings as well as the results on various diabetes datasets.

Lai [13] proposed a predictive model for diabetes mellitus using two machine learning techniques. This objective of the study was achieved using recent records of 13,309 Canadian patients from the age of 18 to 90 years, including bio data like body mass index, sex, triglycerides, blood glucose, age, high-density lipoprotein, blood pressure, and low-density lipoprotein, The predictive model was built using Logistic Regression and Gradient Boosting Machine (GBM) techniques. The results of the study showed that the Receiver Operating Characteristic of the gradient boosting model was 84.7% with a sensitivity of 71.6% while the ROC of the logistic regression model was 84% with a sensitivity of 73.4%. The gradient boosting model and logistic regression models were tested and compared against random forest and decision tree model although the combination of GBM and LR had a higher performance.

Soni [14] In their work titled "Diabetes Prediction using Machine Learning Techniques" proposed the use of Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. The algorithms considered for the model were K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). The result of the experiment showed that Random Forest achieved the most optimal accuracy compared to the other techniques.

III. METHODS AND METHODOLOGY

In this section, the dataset, materials and methods used in this study are discussed as well as the evaluation matrices of the system. The flow of the system is illustrated in Fig 2 which shows the architectural diagram of the diabetes prediction model. From the illustration, there are 5 modules which make up the model and they include:

i. Dataset Collection
ii. Data Pre-processing
iii. Clustering
iv. Build Module
v. Evaluation

The modules are briefly discussed below.

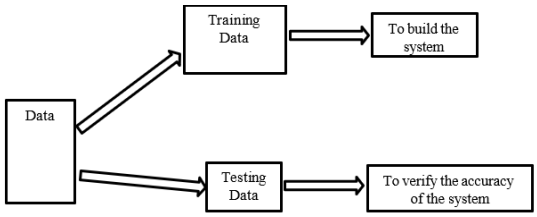

Fig 2: Diabetes Prediction Model

Table 1: Dataset Description of Proposed Machine Learning Algorithm

Features	**Description**
Pregnancies	The condition of being pregnant
Glucose	Simple sugar (monosaccharide)
Blood Pressure	The force of blood pushing against the wall of the arteries
Skin Thickness	Skin thickness is determined by collagen which increases in insulin-dependent diabetes mellitus (IDDM)
Insulin	The polypeptide hormone that regulates carbohydrate metabolism
BMI	Body Mass Index is a person's weight in kilograms divided by square height in meters.
Diabetes Pedigree Function	A function which scores likelihood of diabetes based on family
Age	Patient age

A. Data Collection

The Diabetes dataset from [15] has been applied for the development of the system. The dataset has a dimension of 20001 rows x 9 columns and consists of 2000 rows of diabetes data and 8 attributes. The data was considered to highlight eight different attributes contained in the chosen dataset, the attributes include: age, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function and age. The selected attributes were considered appropriate for determining the outcome. All the attributes have been expressed in numerical values. All the chosen attributes can be used to analyze a patient's condition, determining and diagnosing diabetes in a patient using the machine learning system.

B. Data Pre-processing

The raw Kaggle dataset was already structured in such a way that most of its information and attributes were well organized and stored appropriately. However, some irrelevant attributes were removed, missing values were also manually inputted as attributes such age, blood pressure and BMI cannot have a negative value or a value of zero. Finally, the entire data was scaled to normalize all the values.

C. Clustering

Clustering is a term in unsupervised learning where patterns are grouped into classes [16]. In the clustering method, the categorization of patterns is carried out by grouping patterns with similarities in a cluster whose members are more like each other than to patterns of other clusters [17]. In the case of the selected dataset, all records were properly grouped into diabetic and non-diabetic after which a class label of 0 and 1 was achieved for each record.

D. Build Module

The build module is regarded as the most important phase in the process as it includes the model building which is used for prediction. This study has implemented various machine learning algorithms for predicting diabetes which include SVM, logistic regression, decision tree, random forest and KNN.

E. Evaluation

The evaluation phase is the final step of a prediction model [18]. In the case of this study, overall accuracy and confusion matrix were used in the evaluation of the prediction results. Processing time was also used as a performance metric as it provides a unique perspective in comparative analysis.

i. Overall Accuracy

This is the average of the sensitivity and specificity of a test [19]. Therefore, the Overall Accuracy is the share of the correctly categorized instances. This metric is one of the most widely used.

Overall Accuracy =
TN+TPTP+FP+FN+TN Where TN =
True Negative,

TP = True Positive,

FN =False Negative and FP = False Positive.

ii. True Positive Rate (TPR)

True positive (TP) can be defined as the percentage of certain cases accurately identified [20]. True positive rate or Sensitivity (SN) can be computed as the amount of accurate positive predictions divided by the number of positives [21]. We have the greatest sensitivity as 1.0 and the worst is 0.0. Below is the mathematical definition: True Positive =

TPTP+FN

iii. True Negative Rate (TNR)

TNR is the percentage of negative instances correctly classified. It is also known as Specificity (SP) and it is computed as the amount of accurate negative predictions split by the total number of negatives. Just as in Sensitivity, the best Specificity is 1.0 where the worst is 0.0. Below is the mathematical definition: True Negative = TNTN+FP.

iv. The Area Under the Curve (AUC)

AUC is a measure of ranking performance. It is also used to measure classification performance, gathering over decision thresholds as well as class

and cost skews. AUC signifies the accuracy of a classifier. A large area is an advantage to the classifier.

IV. RESUTS AND ANALYSIS

The performance, model predictions, analysis, and results are discussed in this section.

A. SVM

Figure 12 shows the results from the SVM model. The accuracy rate achieved by SVM is 79%, the precision is 73% while the recall is 0.6.

B. Logistic Regression

The logistic regression classifier achieved an accuracy of 78%, the precision is 73% while the recall is 0.6.

C. Decision Tree

The decision tree had the second-best accuracy at 81%. The precision is 77% while the model's recall is 0.63.

D. Random Forest

The best performing classifier is random forest with an accuracy of 95%, the precision was also the highest with 94% and a recall of 0.91.

E. KNN

K-Neighbors had an accuracy of 80% and a precision of 74% and a recall of 0.63.

F. Figures and Tables

The figures below are the summary of the findings and results achieved from the comparative analysis.

Table 2: Various Algorithm Results

Algorithm	Accuracy	Precision	Recall	Processing Time
SVM	79%	73%	0.6	7s
LR	78%	75%	0.53	4s
DT	81%	77%	0.63	5s
RF	95%	94%	0.91	5s
KNN	80%	74%	0.63	7s

The result from the comparative evaluation shows that logistic regression has the lowest accuracy with a percentage of 78%, it also has a precision of 75%. SVM also had a low accuracy rate of 79%. However, it had a precision 0.6. K-Neighbors had a good accuracy score of 80% however had a precision of 74%. Decision tree was the second-best classification model with 81% and a precision of 77%. Finally random forest was the best performing algorithm with a performance of 95%, it had a precision of 94% and a recall of 0.91.

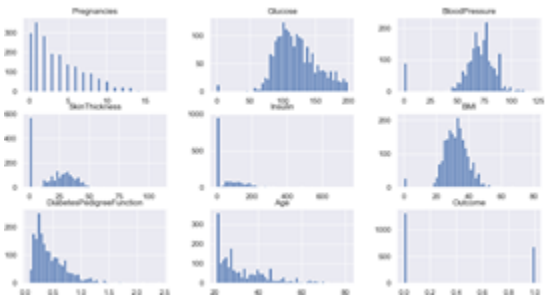

Fig 3: Histogram of Diabetes Dataset

Fig 4: Histogram of Age Attribute

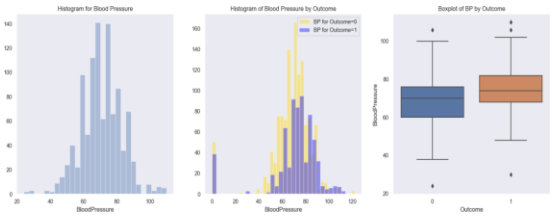

Fig 5: Model Histogram of Attribute Blood Pressure

Fig 6: Histogram of Attribute BMI

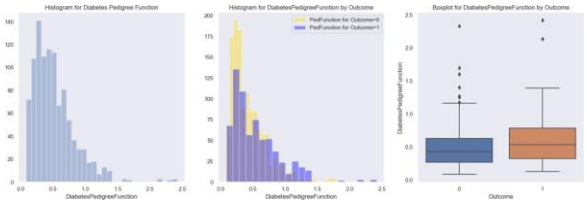

Fig 7: Histogram of Attribute Diabetes Pedigree Function

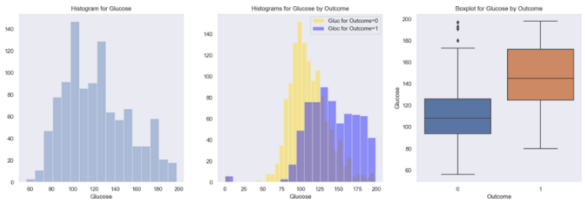

Fig 8: Histogram of Attribute Glucose

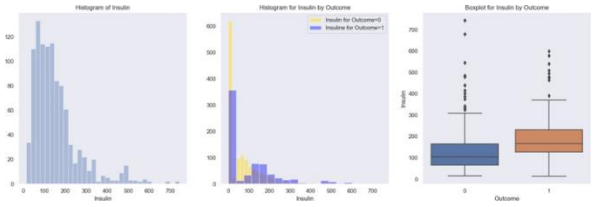

Fig 9: Histogram of Attribute Insulin

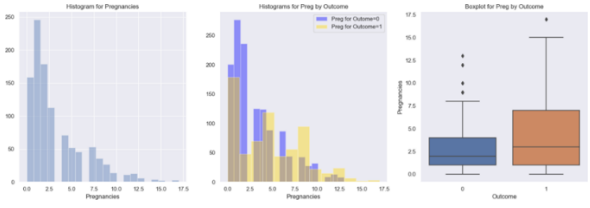

Fig 10: Histogram of Attribute Pregnancies

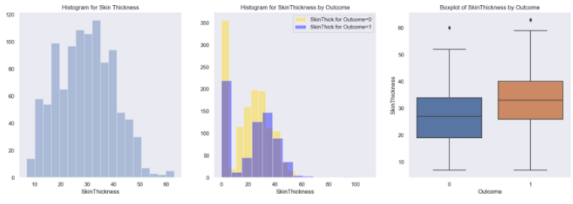

Fig 11: Histogram of Attribute Skin Thickness

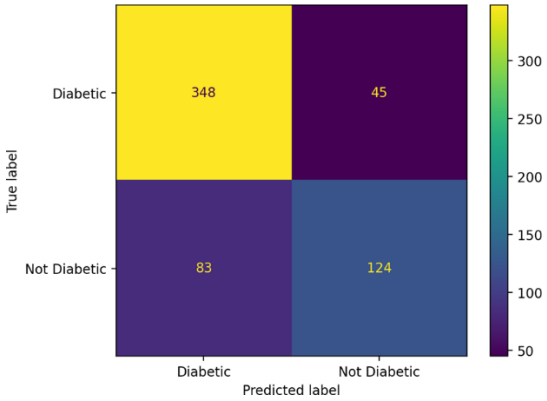

Fig 12: SVM Confusion Matrix

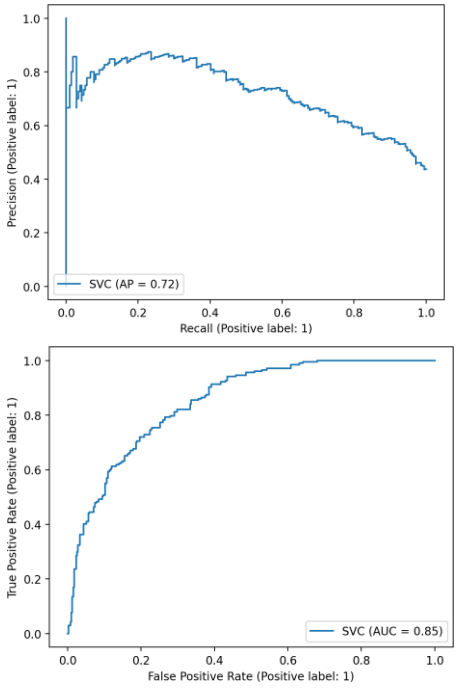

Fig 13: AP and AUC Result of SVM Model

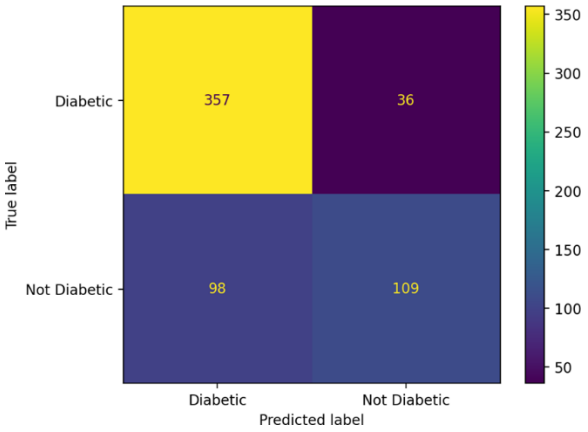

Fig 14: LR Confusion Matrix

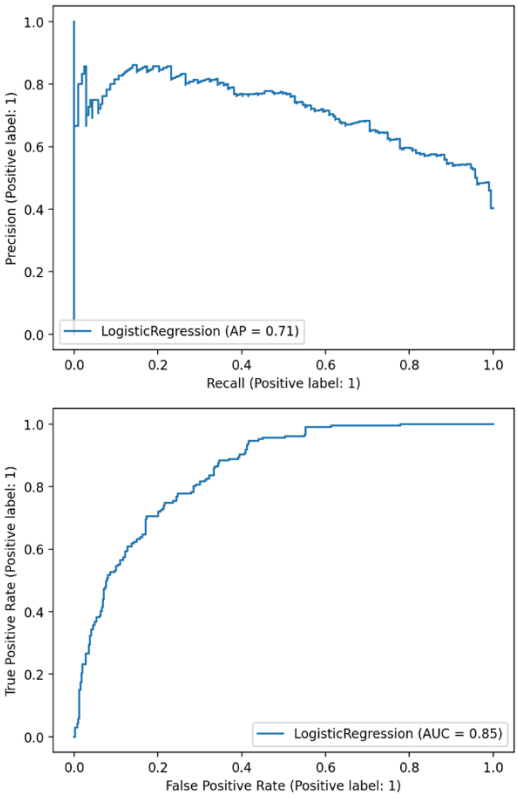

Fig 15: AP and AUC Result of LR Model

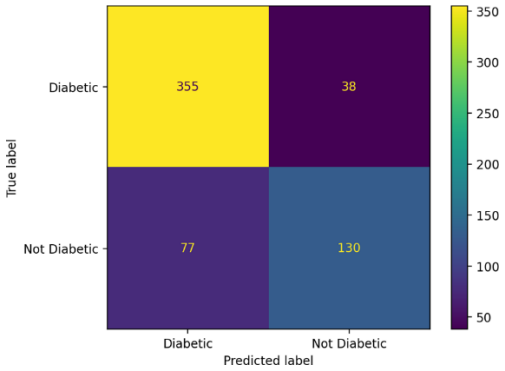

Fig 16: DT Confusion Matrix

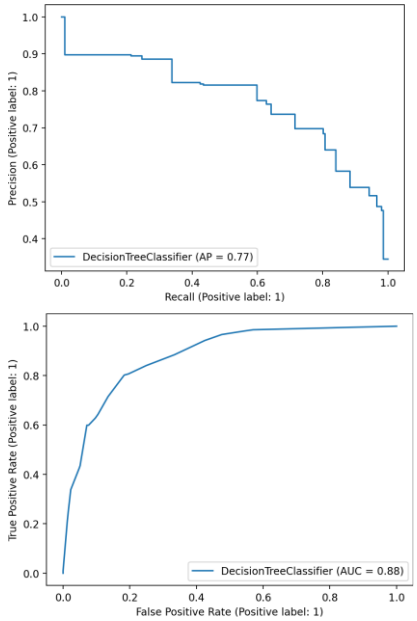

Fig 17: AP and AUC Result of DT Model

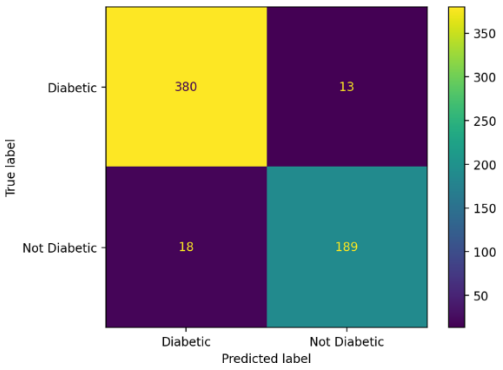

Fig 18: RF Confusion Matrix

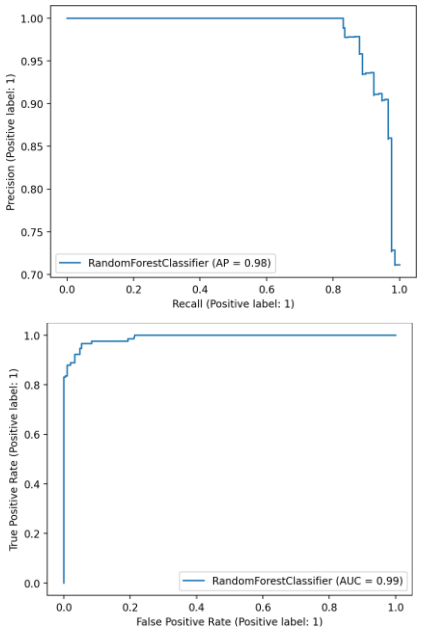

Fig 19: AP and AUC Result of RF Model

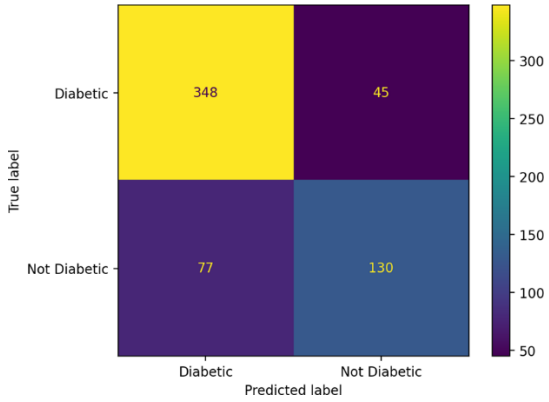

Fig 20: KNN Confusion Matrix

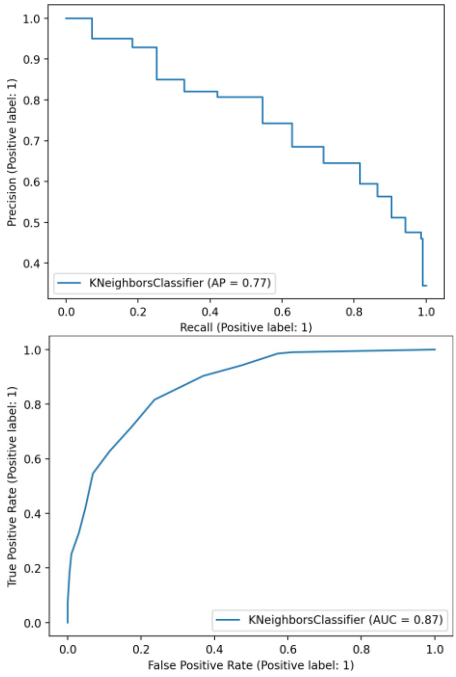

Fig 21: AP and AUC Result of KNN Model

CONCLUSION

In conclusion, this study made use of different machine classification techniques to build a model to predict diabetes. Overall accuracy performance metrics were then applied in other to evaluate the level of accuracy of the models created with showed that the best performing model is random forest with an accuracy of 95%. The development of the model was based on variables obtained from the dataset which was acquired from an open source. SVM, logistic regression, decision tree, random forest and KNN algorithms were applied in this study. A limitation of this system is that it cannot solve all the metabolic health problems therefore the system has to be limited to diabetes. However, in the future, the system models could be improved and repurposed to be used for detecting other types of metabolic illnesses.

REFERENCES

[1] Bohr, Adam, and Kaveh Memarzadeh. "The Rise of Artificial Intelligence in Healthcare Applications." Artificial Intelligence in Healthcare, June 2020, pp. 25–60, https://doi.org/10.1016/B978-0-12-818438-7.00002-2.

[2] Davenport, Thomas, and Ravi Kalakota. "The Potential for Artificial Intelligence in Healthcare." Future Healthcare Journal, vol. 6, no. 2, June 2019, pp. 94–98, https://doi.org/10.7861/futurehosp.6-2-94.

[3] Girasa, Rosario. "AI as a Disruptive Technology." Springer Link, edited by Rosario Girasa, Springer International Publishing, 2020, pp. 3–21, link.springer.com/chapter/10.1007%2F978-3-030-35975-1_1.

[4] Whelan, Rob. "Understanding Data Science, Artificial Intelligence, and Machine Learning." 2nd Watch, 13 Jan. 2021, www.2ndwatch.com/blog/understanding-basics-data-science- artificial-intelligence-machine-learning/.

[5] Hu, F. B. "Globalization of Diabetes: The Role of Diet, Lifestyle, and Genes." Diabetes Care, vol. 34, no. 6, May 2011, pp. 1249–57, https://doi.org/10.2337/dc11-0442.

[6] Sivarajah, Uthayasankar, et al. "Critical Analysis of Big Data Challenges and Analytical Methods." Journal of Business Research, vol. 70, Jan. 2017, pp. 263–86, https://doi.org/10.1016/j.jbusres.2016.08.001.

[7] Kazzazi, Fawz. "The Automation of Doctors and Machines: A Classification for AI in Medicine (ADAM Framework)." Future Healthcare Journal, vol. 8, no. 2, May 2021, pp.e257–62, https://doi.org/10.7861/fhj.2020-0189.

[8] Gliklich, RE, et al. IEEE Standard for an Architectural Framework for the Internet of Things (IoT). New York, Usa Ieee, 2020.

[9] Arentze, T. A. "Spatial Data Mining, Cluster and Pattern Recognition." International Encyclopedia of Human Geography, 2009, pp. 325–31, https://doi.org/10.1016/b978-008044910-4.00524-1.

[10] Dash, Sabyasachi, et al. "Big Data in Healthcare: Management, Analysis and Future Prospects." Journal of Big Data, vol. 6, no. 1, June 2019, https://doi.org/10.1186/s40537-019-0217-0.

[11] Koh, Hian Chye, and Gerald Tan. "Data Mining Applications in Healthcare." Journal of Healthcare Information Management: JHIM, vol. 19, no. 2, 2005, pp. 64–72, pubmed.ncbi.nlm.nih.gov/15869215/

[12] Palanisamy, Venketesh, and Ramkumar Thirunavukarasu. "Implications of Big Data Analytics in Developing Healthcare Frameworks – a Review." Journal of King Saud University - Computer and Information Sciences, vol. 31, no. 4, Oct. 2019, pp. 415–25, https://doi.org/10.1016/j.jksuci.2017.12.007.

[13] Lai, Hang, et al. "Predictive Models for Diabetes Mellitus Using Machine Learning Techniques." BMC Endocrine Disorders, vol. 19, no. 1, Oct. 2019, https://doi.org/10.1186/s12902-019-0436-6.

[14] Soni, Ankit Narendrakumar. "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques." SSRN Electronic Journal, vol. 9, no. 9, 2020, https://doi.org/10.2139/ssrn.3642877.

[15] ukani. "Diabetes Data Set." Kaggle.com, 2020, www.kaggle.com/vikasukani/diabetes-data-set.

[16] Kaushik, Saurav. "An Introduction to Clustering & Different Methods of Clustering." Analytics Vidhya, 11 Mar. 2019, www.analyticsvidhya.com/blog/2016/11/an-introduction-to- clustering-and-different-methods-of-clustering/.

[17] Bandyopadhyay, Sanghamitra. Unsupervised Classification : Similarity Measures, Classical and Metaheuristic Approaches, and Applications. Springer, 2013.

[18] Seliya, Naeem, et al. "A Study on the Relationships of Classifier Performance Metrics." 2009 21st IEEE International Conference on Tools with Artificial Intelligence, Nov. 2009, https://doi.org/10.1109/ictai.2009.25.

[19] Alberg, Anthony J., et al. "The Use of 'Overall Accuracy' to Evaluate the Validity of Screening or Diagnostic Tests." Journal of General Internal Medicine, vol. 19, no. 5, May 2004, pp. 460–65, https://doi.org/10.1111/j.1525-1497.2004.30091.x.

[20] Armah, Gabriel Kofi, et al. "A Deep Analysis of the Precision Formula for Imbalanced Class Distribution." International Journal of Machine Learning and Computing, vol. 4, no. 5, 2014, pp. 417–22, https://doi.org/10.7763/ijmlc.2014.v4.447.

[21] Baratloo, Alireza, et al. "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity." Emergency, vol. 3, no. 2, 2015, pp. 48–49, www.ncbi.nlm.nih.gov/pmc/articles/PMC4614595/.