On The Survival Assessment of Diabetic Patients Using Machine Learning Techniques

Adeboye, Nureni Olawale (PhD)[1*], Adesanya, Kehinde Kazeem[2]

[1]*Department of Mathematics & Statistics, Federal Polytechnic, Ilaro, Ogun State, Nigeria*

[2]*Department of Health Information Management, Ogun State College of Health Technology, Ilese ijebu Ode, Ogun state Nigeria.*

Abstract: **The extraordinary improvement in biotech and medical sciences have given rise to an impactful data production from stour Electronic Health Records (EHRs), and it has contributed significantly to the Kaggle source from which the data for this research was obtained. The dataset consists of 1416 recorded cases of diabetic patients from 130 various hospitals in the United States. This study thus assesses the survival rate of diabetic patients using machine learning techniques, and determined the duration it will take a diabetic patient to survive based on the application of the most appropriate algorithm. The research tested the application of four different algorithms which include support vector machine, logistic regression, decision tree and k-nearest neighbors' algorithm. In line with their accuracy measured by f1-score, precision, recall and support metrics; k-nearest neighbors is seen to outperform all other algorithms for predicting the survival rate of the patients. The research also revealed that it takes a diabetic patient 30 days to survive if the patient is placed on medications according to the available information, and that the medication given to the diabetic patients is less effective in the aged patients and more effective among the younger patients.**

Keywords**: Accuracy, Algorithms, Diabetic Patients, Machine Learning, Survival rate.**

I. INTRODUCTION

Pertinent researches in the area of technology that uses biological systems and living organisms to develop different outputs results unremittingly in a self-evident and economic data production, thereby heralding the science of biotechnology into the realm of big data. In addition to this lofty performance, there is a myriad of electronic machines from various research fields which culminate into data generation, and these include Super-Resolution Microscopy, Spectrometry Technologies for biomolecules and small molecules, Magnetic Resonance Spectroscopy, just to mention few. Though these technologies produce valuable data, but they do not give researchers insight into the analytical meaning of the generated data. Thus, Knowledge Discovery in Biological data has becomes essential and logically inescapable; the primary aim is mainly to research into the rapidly increasing body of such official data and set the basis for providing genuine responses to fundamental questions in biological and medical sciences ref. [6]. In the hybrid field of biotechnology, Diabetes Mellitus (DM) is one of the mostly diagnosed ailments in the categories of human-threatening and life quality reducing diseases ref. [2].

According to ref. [13], DM is a metabolic disorder in which the amount of sugar in the blood is increased beyond necessary. Insulin deficiency increases the glucose levels in the blood and subdue the metabolism of carbohydrates, fat and proteins. It is the most normal endocrine issue, affecting more than 415 million individuals in the entire world. Diabetes development is emphatically connected to hormonal and metabolic issues, brought about by constant hyperglycemia. Diabetes covers a wide scope of heterogeneous pathophysiological conditions. Difficulties like harmed nerves, eyes, kidneys, and different organs might emerge when high glucose from diabetes is not treated on schedule.

According to [20], diabetes is categorized into two major clinical types according to the etiopathology of the disorder, which are Type 1 diabetes (T1D) and Type 2 diabetes (T2D). 90% of all diabetic patients are known to be suffering from T2D and thus regarded as the most common form of diabetes, and [21] emphasized that the increasing burden of T2D has become a major concern in healthcare management. T2D is mainly characterized by insulin resistance and the main causes include but not limited to poor Medicare, dietary habits and heredity. T1D on the other hand, is opinioned to be due to auto immunological destruction caused by a chronic condition in which the pancreas produces little or no insulin. T1D usually manifest in adolescence and it has been established to affects almost 10% of all diabetic patients globally, resulting in symptoms such as increased thirst, frequent urination, fatigue, blurred vision and hunger. Other classifications of DM on the basis of insulin secretion profile include Endocrinopathies, Gestational, Mitochondrial, MODY (Maturity Onset Diabetes of the Young), Neonatal, and Pregnancy diabetes.

According to data published by the Global Burden of Disease (GDB) in 2017, diabetes global burden increased greatly from 11.3 million in 1990 to 22.9 million in 2017 with a 102% increase. Due to high diabetes death rate and indisposition as well as related disorders, prevention and treatment attracts broad and significant interest. Insulin is the main treatment for Type1, even though in certain cases insulin is also provided to Type2 diabetic patients, when hyperglycemia cannot be controlled through diet, weight loss, exercise and oral medication. The most common anti-diabetic agents include sulfonylurea, metformin, alpha glucosidase inhibitor, peptide

analogy, non-sulfonylurea, secretagogue, etc. Most of the present anti-diabetic agents, exhibit numerous side-effects. In addition, insulin therapy is related to weight gain and hypoglycemic events. According to ref. [14], the diagnoses of diabetes largely depends on the level of blood glucose in the patient system. The tolerant level of fasting plasma glucose has been established at 7.0 mml/L.

World Health Organization [20] estimated the number of deaths caused by diabetes in 2019 as 1.5 million; and reports from International Diabetes Federation (IDF, 2019) estimated the global diabetes prevalence to be 9.3% (463 million people). This was projected to rise to 10.2% (578 million) and 10.9% (578 million) by 2030 and 2045 respectively. Ref. [18] opined that almost half a billion people are living with diabetes globally and the projection has been estimated at 25% and 51% respectively in 2030 and 2045.

Diabetes mellitus is a chronic health problem which is devastating, yet preventable consequences. It currently comprises the most noteworthy morbidity and mortality of all prolonged non-transferable sicknesses in Africa. In Nigeria, diabetes represents 3–15% of clinical affirmations in most wellbeing offices. As indicated by ref. [4], People living with type 2 diabetes are more powerless against different types of both short-and long run complications, which regularly lead to their unexpected death. This assessed number of deaths is comparative in size to the collective deaths experienced from many communicable diseases that are of major public health priority.

Due to its high mortality rate, necessary bio-statistics measures are really needful in other to estimate the duration of survival of a diabetic patient. However, previous studies have not been able to accurately estimate the survival rate attributed to both in and out patient of diabetes mellitus. This study therefore is concerned with the adoption of machine learning techniques in accurately determining the survival rate of diabetes patients.

Survival Analysis deals with the application of methods to estimate the likelihood of a demographic event (death, survival, decay, child-birth etc.) occurring over a variable time period. The traditional statistical methods applied in the area of survival analysis include the Kaplan Meier (KM) estimator curve and the Cox-proportional hazard (PH) models ref. [9], as adopted in ref. [11].

The application of deep learning techniques in biometrics analysis is a useful attempt to accomplishing available largeish diabetes-related data for knowledge extraction. The extreme social implication of diabetes occasioned large data extraction as a germane priority in biometrics, which undoubtedly has been of tremendous assistance in medical research. Without prejudice, deep learning and classification algorithms techniques had been of great assistance in the diagnosis health related issues during clinical trials ref. 5]. Among the authors who have employed machine learning techniques in the study of diseases are ref. [17, 10, 14, 19, 8, 15, 16, 1,12].

Ref. [17] employed Linear Support Vector Machine (LSVM) algorithm in the classification and prediction of diabetes disease with a metric of 75.5 % accuracy. Ref. [10] developed a predictive model for renal graft status and survival period using the Byes' Net Classifier on the data collected from the University of Toledo Medical Center. According to ref. [14], the data of Surveillance, Epidemiology and End Results (SEER) Program was employed to developed a predictive model for the classification of the survival of lung cancer patients. Predictive model was formulated using different decision trees algorithms, of which the algorithms used had accuracies of 73.61%, 74.45%, 76.80%, 85.45% and 91.35% respectively for the 6 months, 9 months, 1 year, 2 years and 5 years' survival dataset. Ref. [19] developed a predictive model for the classification of diabetes disease using support vector machine (SVM)s. The data contained 500 and 268 cases of patients that did not survived and those that survived respectively. SVM classifier was employed to train the predictive model using the 10-fold cross validation exercise. The empirical results showed that the SVM had an accuracy of 78% with true positive and true negative values of 80% and 77% respectively. Ref. [8] developed a predictive model for diabetes diagnosis using the fuzzy c-means clustering and support vector machines. The study used the techniques to formulate the predictive model for the diagnosis of diabetes and the results showed that the fuzzy means clustering algorithm outperformed the SVM algorithm with an accuracy of 94.3% alongside a true positive rate of 95.4%. Ref. [17] used clinical variables to developed a predictive model for the survival of pediatric sickle cell disease (SCD). The authors employed the use of fuzzy logic based model using three (3) clinical variables. [18] uses naïve Bayes' classifier in a supervised machine learning algorithm to predict the survival of pediatric HIV/AIDS patients. The 10-fold cross validation training technique was used to train the predictive model for survival classification and the results showed that the classifier was able to predict the survival of patients with an accuracy of 68%. Ref. [1] carried out comparative study on different machine learning models, such as the decision tree classification, K-Nearest Neighbor (KNN), Linear Support Vector Machine (LSVM) and Naive Bayes. The metrics used in comparing the performance of the classification models on prediction of the diabetes disease were, the Accuracy, recall and precision. The result of the study showed that LSVM has better performance in classification of the diabetes dataset collected from a medical center in Bangladesh. Ref. [12] engages the comparison of two boosting algorithms in terms of efficiency with four (4) other single based classifiers on cardiovascular official data. Gridsearch of 5fold cross validation was carried out using multiple hyperparameters for each model, and the research confirmed that boosting algorithms are better predictors compared to single based classifiers. Ref. [21] used machine learning techniques to determine the predictive estimates of biomarkers in the

development of type 2 diabetes mellitus (T2DM). Their results suggested that inflammatory biomarkers and HOMA-IR have a strong prognostic value in predicting progression to T2DM and ascertained that Machine learning techniques provided more accurate results to better understand the implications of these features in terms of progression to T2DM. They opined that a successful therapeutic approach based on inflammatory biomarkers and HOMA-IR can avoid progression to T2DM and thus improve long-term survival. Ref. [22] predicted the heart failures patients' survival using nine classification models viz-a-viz Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient classifier (SGD), Random Forest (RF), Gradient Boosting classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB) and Support Vector Machine (SVM). Experimental results demonstrated that ETC outperforms other models and achieves 0.9262 accuracy value with SMOTE in prediction of heart patient's survival. Ref. [23] evaluated the clinical applications of body mass index (BMI) and a percussion-entropy-based index (PEI_{NEW}) for predicting the development of diabetic peripheral neuropathy (DPN) in a group of type 2 diabetes mellitus (DM) patients. Kaplan-Meier survival analysis employed showed that the diabetics patients with BMI $>$ 30 had a significantly higher cumulative incidence of Peripheral neuropathy on follow-up than those with BMI $\leqq$ 30.

Thus in this study, Machine Learning Techniques were adopted in accessing the survival of a diabetic patient using secondary data collected from Kaggle Dataset Repository of US hospitals on the reported and recorded cases of diabetics for some years. The purpose of the study is to investigate the performances of different machine learning techniques in the survival analysis of a diabetic patient based on the available factors responsible for how long a patient stays in hospital admission.

II. MATERIALS AND METHODS

2.1 Materials

The data set used for this research was adapted from Kaggle Dataset Repository source (www.kaggle.com/brando). It contains sample of reported cases in 130 Hospitals in United States of America, over a period of ten years (2009-2018). Over this period, 1416 reported cases were observed from the sourced data. The attributes of the dataset are as explained in Table 1.

2.2 Methods

This research considered four different machine learning techniques of logistic regression, support vector machine, k-Nearest neighbor algorithm and decision tree; and subject them to comparison based on information obtainable from four different evaluation metrics of Precision, Recall, F1, and support metrics in order to ascertain the best model for predicting the treatment of a diabetic patient.

2.2.1 Machine Learning Techniques

Logistic Regression

Logistic regression involves modelling of a binary dependent variable via a logistic function. Empirically, a binary logistic model has a dependent variable with two possible values labeled "0" and "1" representing failure and success respectively. In the logistic model, the log-odds for the value labeled "1" is a linear combination of one or more independent variables [24].

Considering a model with two predictors, X_1 and X_2, and one binary response variable Y, denoted as $P=P(Y=1)$. We assume a linear relationship between the predictor variables and the log-odds of the event that $Y = 1$. This linear relationship can be written in the following mathematical form (where ℓ is the log-odds, b is the base of the logarithm, and β_i are the parameters of the model):

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (1)$$

We can recover the odds by exponentiation of the log-odds to have

$$\frac{p}{1-p} = \ell^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \qquad (2)$$

By simple algebraic manipulation, the probability that $Y = 1$ is

$$p = \frac{\ell^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{\ell^{\beta_0+\beta_1 x_1+\beta_2 x_2}+1} = \frac{1}{1+\ell^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

$$= S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \qquad (3)$$

Where S_b is the sigmoid function with base b. Equation (3) shows that once β_i are fixed, the log-odds that $Y=1$ can easily be computed for a given observation.

k-nearest neighbors' algorithm (k-NN)

k-NN is a distribution free technique mostly used for regression tree and classification algorithm, where k is typically a small positive integer. In both situation, the output results are a function of the k-closest training inputs in the feature space. An object is classified by a plurality choice of its neighbors, through the assignment of the object to the most common class among its k nearest neighbors. suppose $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

Given pairs of variables $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ taking values within a defined real-valued space, where Y is the class label of X, so that $X|Y$ is distributed with probability distribution P_r. Let $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \cdots, (X_{(n)}, Y_{(n)})$ be a

reordering of the training data such that $||X_{(1)} - x|| \leq \cdots \leq ||X_{(n)} - x||$.

Decision Tree

The algorithm is among the most popular machine learning algorithms given their intelligibility and simplicity. It uses a tree like structure as a predictive model. The classification trees posited Tree models where the target variable can take a discrete set of values, in which leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The algorithm can equally be used in classification and regression. Decision trees where the target variable can take continuous values are called regression trees**.** Technically, the model is similar to proportional stratification in statistics ref. [25].

Support Vector Machine

The algorithm is a supervised learning model that equally analyze data for classification and regression analysis. The training algorithm builds model that assigns new choice to each category making it a non-probabilistic binary linear classifier. SVM maps training choices to points in space so as to maximize the width of the gap between the two categories, with subsequent mapping of the new selection until the desired results are achieved. In computation, SVM amounts to minimizing an expression of the form

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max(0,1-y_i(w^T x_i - b))\right] + \lambda||w||^2 \qquad (4)$$

2.2.2 Evaluation Metrics

F1-score

The f-score is a measure of test accuracy in statistical analysis of binary classification. Mathematically, it is expressed as

$$f1 = 2.\frac{precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad (5)$$

Where TP denotes number of true positives, FP is number of false positives and FN is number of false negatives.

Precision and Recall

Precision metric is the fraction of relevant instances among the retrieved instances, while recall is the fraction of retrieved relevant instances among all relevant instances. According to [12], Recall helps when the cost of false negatives is high and both metrics are calculated as follows:

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (7)$$

Where TP, FP, FN are true positives, false positives and false negatives respectively.

Support

The support is the number of occurrences of each class of number of true positives.

III. RESULTS AND DISCUSSION

Based on the research focus, the descriptive statistics of the processed data were presented in Figure 1. It shows the bar charts presentation of all the 46 variables available in the information of 1416 patients contained in the data obtained for the study.

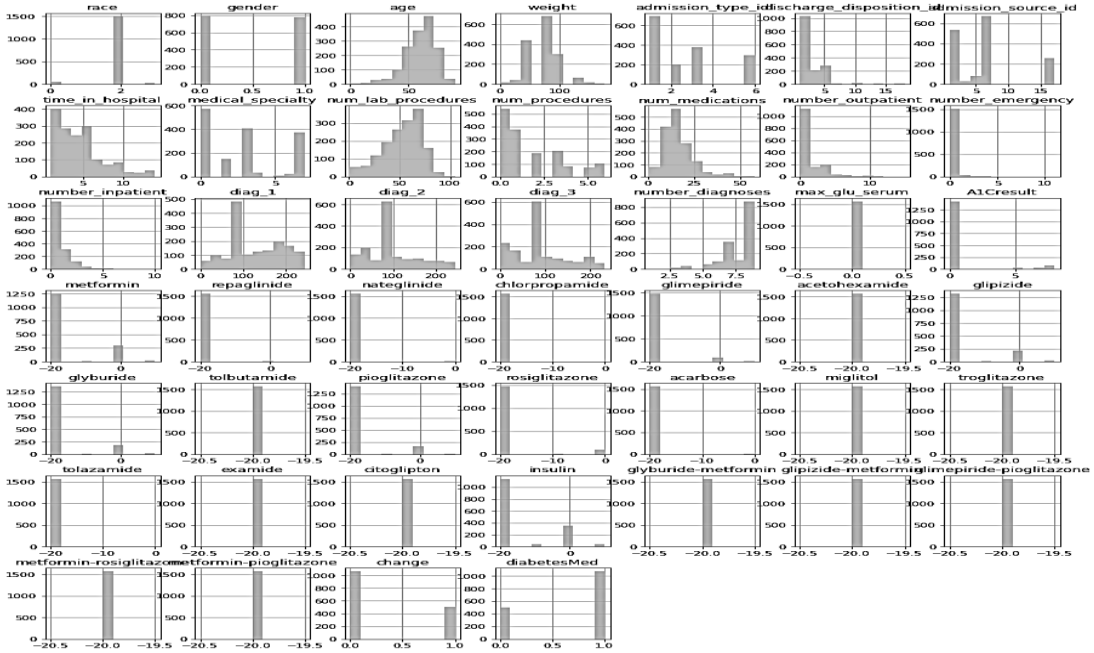


Fig 1: Bar chart presentation of all variables of the diabetic patient under study

3.1 Survival Rate Based on Medication Usage

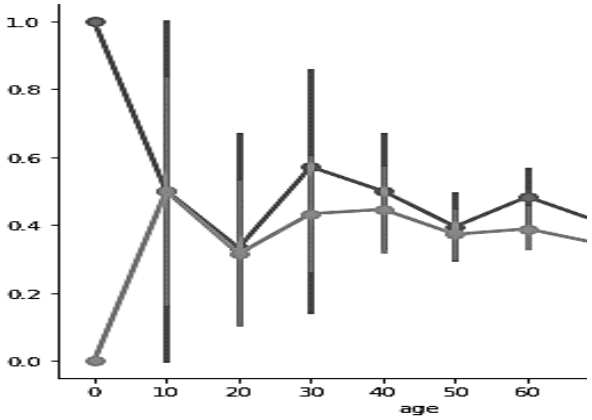


Fig 2: Line graph of patient (age) on medication usage and readmission probability

0 = not on diabetics' medication; 1 = on diabetics' medication

Fig 2 shows the odd of patients (across their age groups) been readmitted within the period of 30 days' treatment. The study revealed that patients below the age of ten who is on medication has a zero probability of been readmitted within 30 days (i.e. no readmission) but if such patient is not on medication, the probability of readmission within the period of 30 days is 1.0. For the patients within the age (10 -20 years), the probability of readmission for both patients on Medicare and those not on Medicare is 0.5. In the case of (20 – 30 years) age group, patients on Medicare has a slightly lower probability of readmission than those patients who are not on Medicare. Patients within the age group of (30 – 40 years) has the probability of 0.45 for patients on Medicare while those not on Medicare have a probability of almost 0.6 of being readmitted, while the age group (40 – 50 years) has a lower probability of readmission for patients on Medicare than those that are not on Medicare. These probabilities can be viewed as almost 0.4 and 0.4 respectively, and the same occurrences can be observed across the remaining age groups. It was observed that patients on medication has a lower probability of readmission than those not on medication, and this shows that the treatment method(s) adopted contribute to the chances of a patient not been readmitted. Also, the potency of the treatment gradually drops as the patients ages increase, as reflected in the upward trend of probability of readmission line.

3.1 Model Comparison

Table 1: Comparison of Marching Learning Models

Metrics	Outcomes	Precision	Recall	F1-score	Support
Logistic Regression					
	Re-admit	0.64	0.73	0.68	184
	Don't Re-admit	0.52	0.42	0.47	130
Accuracy				0.60	314
Macro average		0.58	0.58	0.58	314
Weighted average		0.59	0.60	0.59	314
Support Vector Machine					
	Re-admit	0.59	1.00	0.74	184
	Don't Re-admit	0.00	0.00	0.00	130
Accuracy				0.59	314
Macro avg		0.29	0.50	0.37	314
Weighted avg		0.34	0.59	0.43	314
k-nearest neighbors algorithm (k-NN)					
	Re-admit	0.65	0.74	0.69	184
	Don't Re-admit	0.55	0.45	0.49	130
Accuracy				0.62	314
Macro average		0.60	0.59	0.59	314
Weighted average		0.61	0.62	0.61	314
Decision Trees					
	Re-admit	0.64	0.65	0.66	184
	Don't Re-admit	0.52	0.54	0.53	130

Accuracy				0.60	314
Macro average		0.59	0.59	0.59	314
Weighted average		0.60	0.60	0.60	314

Table 1 presents the comparative results of the considered machine learning techniques efficiency in predicting survival rate of a diabetic patient. From the results, it was observed that K-nearest neighbors' algorithm (k-NN) outperformed the other three (3) techniques based on its accuracy value of 62% measured by F1-Score. This is closely followed by the fi-score of Logistic regression and Decision tree with accuracy of 60%. The least accurate model is the Support Vector Machines which has a value of 59%. The results also showed that K-NN gives the highest precision of 65% and 55% respectively in terms of patients' classification into readmission and non-readmission cases. The precision result of 55% implies that KNN gives the highest chance of survival to any patients undergoing treatments for diabetics.

Fig. 3 compared the adopted models in terms of classification metrics using box and whisker plots as used in [26], for weighted averages of the metrics. These results suggest that both logistic regression and supporting vector machines are perhaps worthy of further examination on this study, closely followed by K-NN. However, the estranged values of weighted precision and F1 score of SVM would not allowed any further consideration of the technique. Logistic regression would have been the best recommended model, but going by the context of this research, KNN outperformed logistic regression model based on its highest rate of precision for the survival of every diabetics' patients.

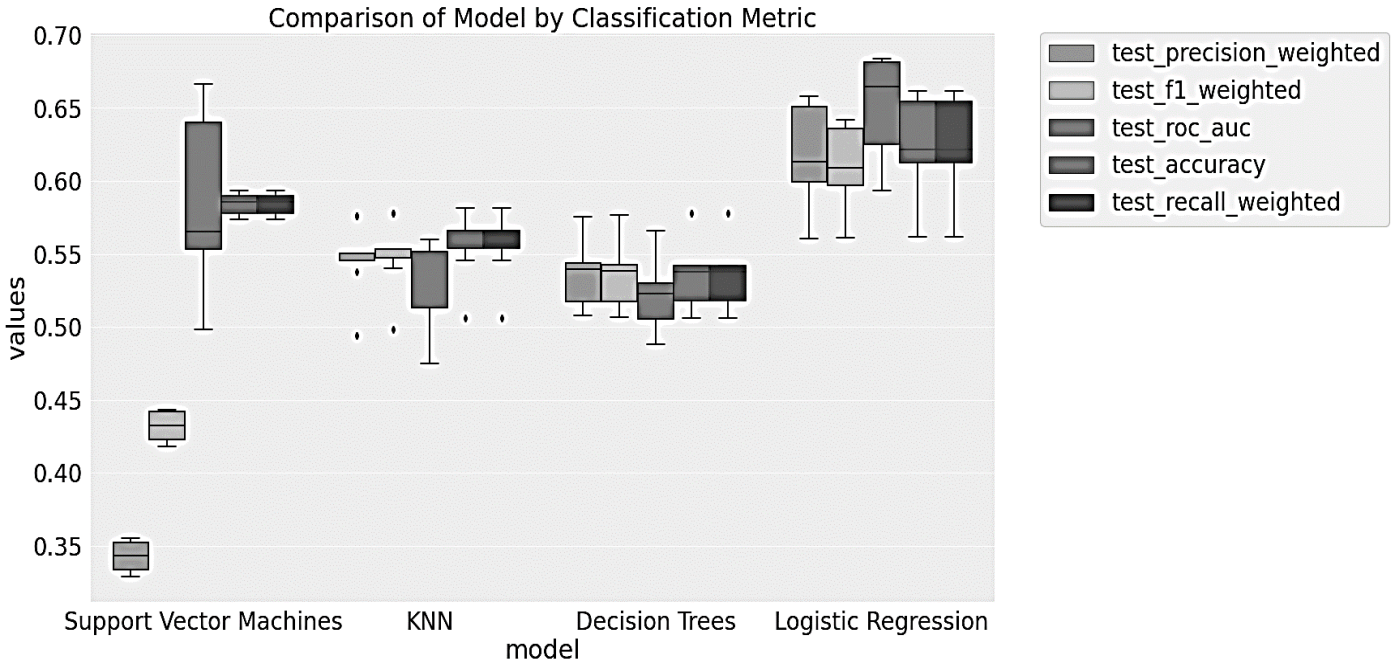


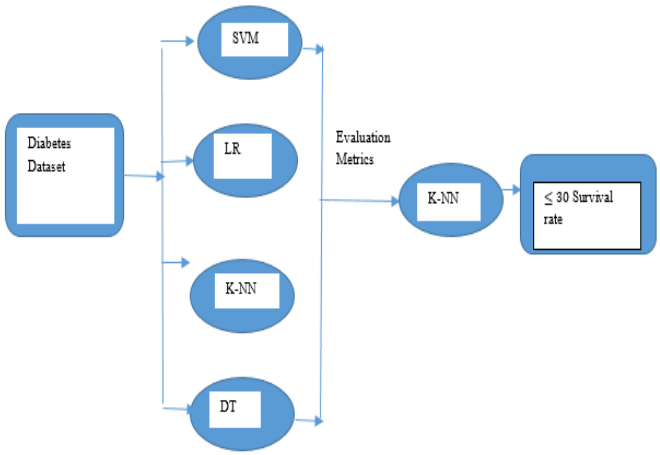
Fig 3: Comparison of model by classification metric

IV. CONCLUSIONS

The study revealed the average duration (in days) it will take a diabetic patient to survive based on readmission into the hospital. Results showed that it will take a minimum of 30 days for a diabetic patient to recover completely if the patient is on medication according to the hospital prescription. In order to generate a suitable survival model for the classification data, the study revealed that the K-nearest neighbors' algorithm (k-NN) is the most preferred algorithm to adopt based on the available information obtained for the study. The research also comes up with possible suggestions and policy recommendations to ensure that the odds of survival in diabetic patient increases. Since the potency of the medication tends to reduce as the age of the patient increases, the study deduced that the treatment method for the aged patient should improve in other to obtain a better odd of survival in diabetic patients.

The research equally recommends a future study where the best adopted algorithm of K-NN is compared with the traditional methods of survival analysis such as Kaplan Meier technique.

Graphical Abstract



ACKNOWLEDGEMENTS

The authors wish to acknowledge the online data producers through which the data for this research was sourced.

REFERENCES

[1]. A. Aladekomo. "The major issues of big data". Nature;498(2018)7453: 255–60, http://dx.doi.org/10.1038/498255a.
[2]. A. Brando. "Diabetics Research on patient survival in 13 US hospitals" https:www.kaggle.com/brando (2017). https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.kaggle.com/brandao/diabetes&ved=2ahUKEwir8fuunvHvAhVxQRUIHU5hBRYQFjAAegQIBBAC&usg=AOvVaw1dPPwCMsZMxXGcvSp4mH_E&cshid=1617973883313
[3]. A. M. Karahoca & T. Alper. "Dosage planning for type 2 diabetes mellitus patients using indexing HDMR". Expert System Application ;39(2012)8 :7207–15.
[4]. A. B. Olokoba, O. A. Obateru, L. B. Olokoba, "Type 2 Diabetes Mellitus: A Review of Current Trends". Oman Medical Journal, 27(2012)4, 269-273. doi: 10.5001/omj.2012.68
[5]. A. J. Lee, B. Ku, J. Nam & D. D. Pham. "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes". IEEE J Biomed Health Inform Mar 2014;18(2013):555–61. http://dx.doi.org/10.1109/JBHI.2013
[6]. A. Worachartcheewan, C. Nantasenamat, P. Prasertsrithong, J. Amranan, T. Monnor, T. Chaisatit. "Machine learning approaches for discerning intercorrelation of hematological parameters and glucose level for identification of diabetes mellitus". EXCLI Journal 12(2015):885–93 [eCollection,2013].
[7]. IDF (2019). Results from the International Diabetes Federation, Diabetes Atlas Committee, 9th Edition.
[8]. J. Jin, H. Min, S. J. Kim, S. Oh, K. Kim, H. G. Yu & A. Veena. "Development of diagnostic Biomarkers for detecting diabetic retinopathy at early stages using quantitative Proteomics". Diabetes Research (2016) 6571976. http://dx.doi.org/10.1155/2016/.
[9]. K. Kaplan, S. Howard & J. Agasa. "Applications of Machine Learning in Cancer Prediction and Prognosis". Cancer Informatics 2(1958): 59-75.
[10]. K. J. Li, J. Y. Kim, S. Dearden. "Mining recent temporal patterns for event detection in time series data on medications". KDD; (2012) :280–8.
[11]. N. O. Adeboye, I. A. Ajibode & O. L. Aako. "On the Survival Assessment of Asthmatic Patients Using Parametric and Semi-Parametric Survival Models. Occupational Diseases and Environmental Medicine, 8 (2020) 2. 50-63. https://doi.org/10.4236/odem.2020.82004.
[12]. N. O. Adeboye & O. V. Abimbola. "An overview of cardiovascular disease infection: A comparative analysis of boosting algorithms and some single based classifiers. Statistical Journal of the IAOS 36 (2020), 1189–1198, DOI 10.3233/SJI-190609.
[13]. P. S. Bradley. "Implications of big data analytics on population health management". Big Data Sep 2013;1(2013),3:152–9. http://dx.doi.org/10.1089/big.2013.0019 [Epub 2013 Sep 5].
[14]. M. A. B Khan, M. J. Hashim, J. King, R. Govenda, H. Mustafa, J. Alkaabi. "Epidemiology of Type 2 Diabetes". Global Burden of Disease and Forecasted Trends (2019). DOI:10.2991/jegh.k.191028.001
[15]. R. Agrawal & R. Srikant. "Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Databases; 2004. p. Marx V. Biology: the big challenges of big data. Nature Jun 13 2013;498(2013)7453: 255–60. http://dx.doi.org/10.1038/498255a.
[16]. P. A. Idowu, O. Agbelusi, & T. A. Aladekomo, " Mining recent temporal patterns for event detection in multivariate time series data". KDD (2017): 280–8.
[17]. P. A. Idowu, O. Agbelusi, & T. A. Aladekomo. The Prediction of Pediatric HIV/AIDS Patients' Survival: A Data Mining Approach. Asian Journal of Computer and Information Systems 4 (2018)3: 87-94.
[18]. S. DuBrava, J. Mardekian, A. Sadosky, E. Jay Bienen, B. Parsons, M. Hopps, J. Markman. "Using random forest models to identify correlates of a diabetic peripheral neuropathy diagnosis from electronic health record data". Pain Med. 18 (2017) 1:107-115. Doi: 10.1093/pm/pnw096.
[19]. S. Pouya, P. Inga, S. Paraskevi, M. Belma, K. Suvi, K., Nigel, U., Stephen, C., Leonor, G., Ayesha, A. M., Katherine, O., Jonathan, E. S., Dominic, B., Rhys, W. (2019). Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045. Diabetes Research and Clinical Practice. 157 (2019), 107843. https://doi.org/10.1016/j.diabres.2019.107843.
[20]. V. Bijalwan, K. Vinay, P. Kumari & P. Jordan. "KNN Based Machine Learning Approach for Text and Documents Mining". International Journal of data base theory and application, 7(2014)1: 61-70.
[21]. WHO (2020). Diabetes: Key Facts. Available fromhttps://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed may 29, 2020].
[22]. R. Garcia-Carretero, L. Vigil-Medina, O. Barquero-Perez (2021). The use of Machine Learning Techniques to Determine the Predictive Value of Inflammatory Biomarkers in the Development of Type 2 Diabetes Mellitus. Metabolic Syndrome and Related Disorders. 19(4); 240-248. .http://doi.org/10.1089/met.2020.0139
[23]. A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. 2020 IEEE International Conference for Innovation in Technology (INOCON) IEEE Access, vol. 9, pp. 39707-39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
[24]. M. Xiao, C. Lu, Na Ta, H. Wei, B. Haryadi, H. Wu (2021). Machine learning prediction of future peripheral neuropathy in type 2 diabetics with percussion entropy and body mass indices. Biocybernetics and Biomedical Engineering, 41 (3), 1140-1149, ISSN 0208-5216, https://doi.org/10.1016/j.bbe.2021.08.001.
[25]. I. Rallis, I. Markoulidakis, I. Georgoulas, G. Kopsiaftis (2020). "A novel classification method for customer experience survey analysis". Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments. 68(2020), 1–9. https://doi.org/10.1145/3389189.3397999.
[26]. J. R. Quinlan (1986). Induction of decision trees. Machine Learning. 1(1):81–106.
[27]. I. D. Dinov (2018). Data Science and Predictive Analytics. Biomedical and Health Applications using R. Doi: 10.1007/978-3-319-72347-1