Survival Model for Diabetes Mellitus Patient Receiving Treatment

AMOO A. O.[1], OYEGOKE T. O.[1], BALOGUN J. A.[1], BAMIDELE S. A.[2], IDOWU P. A[1].
[1]Department of Computer Science and Engineering
Obafemi Awolowo University, Ile-Ife
NIGERIA
[2]Department of Computer Science
Tai Solarin University of Education, Ijebu-Ode
NIGERIA
[1]olawunmikemi@yahoo.com; [1]temitayooyegoke@yahoo.com; [1]jeremiahbalogun@gmail.com;
[2]simirado@gmail.com; [1]paidowu@oauife.edu.ng

Abstract

This study focused on the development of a predictive model for assessing the survival of diabetes mellitus (DM). The study identified the variables monitored during the treatment of diabetes mellitus patients, formulated, simulated and validated the predictive model for the survival time of diabetes mellitus patients. Following the identification of relevant variables, data was collected from 29 patients alongside their survival time. The predictive model for assessing the survival time of DM was formulated using the support vector machines (SVM) and multi-layer perceptron (MLP) classifiers. The models were simulated using the 10-fold cross validation technique via the WEKA® simulation software. The model were then validated and compared based on the mean absolute error (MAE) and the mean square error (MSE) rates. The results showed that 21 factors were assessed among the collected data required for assessing survival time and the MLP showed the better ability to assess the survival time with error rates of 0.000 however had a longer model build time of 0.77 seconds compared to the build time of 0.02 seconds using SVM. The study concluded that information about the assessment of the survival time of patients with DM can provide decision support aimed at providing alternative or continuous treatment to DM patients.

Keywords: Diabetes Mellitus, Predictive Model, Supervised Machine Learning, SVM, MLP

1. Introduction

Communicable diseases have always been the major cause of death around the world however growth in medical research and improvements of life conditions in industrialized countries have brought about a transition from communicable diseases to non-communicable diseases. It is projected that non-communicable diseases trends are likely to exceed communicable diseases in developing nations thus culminating in double burden [1]. Diabetes is a degenerative illness that requires lifelong care to control blood glucose levels and prevent complications. Diabetes sufferers are at risk for developing long-lasting complications that may lead to loss of vision, kidney disease, nerve damage, peripheral circulatory disorders and other serious complications. Also, Diabetic patients are at danger of having stroke and heart disease which can finally lead to all variety of incapacitating and life-endangering complications.

According to the World Health Organization [2], an estimation of 1.6 million deaths were directly caused by diabetes and it was recorded as the seventh leading cause of death worldwide. It was also projected that by 2040, the growing morbidity of the world's diabetic patients will reach 642 million. There is an imperative importance to give attention to these alarming projected cases. In Nigeria, diabetes accounts for 3 to 15 percent of medical admissions into most health facilities [3]. A report by the International Diabetes Federation [4], stated that close to half of deaths due to diabetes are in people under the age of 60 years. Also, about 5.1 million people aged between 20 and 79 years died from diabetes in 2013, accounting for 8.4% of global all-cause mortality among people in these age groups [5]. This estimated number of deaths is similar in magnitude to the combined deaths from several infectious diseases that are major public health priorities, and is equivalent to one death every six seconds.

The recent availability of the electronic health records offers an unprecedented opportunity to apply

predictive analytics to improve the practice of medicine and to infer potentially novel risk factors. The application of data mining can help physicians in various ways such as given meanings to different diagnostic tests, combining information from several sources (sample movies, images, clinical data, proteomics and scientific knowledge), given that support for differential diagnosis and providing patient-specific prediction. Data mining has a great potential to enable healthcare systems to use data more efficiently and effectively thereby reducing the likely costs associated with making decisions [6].

Survival Analysis is one of the most popular methods of data mining which deals with the estimation of the time to an event such as: death, child-birth, radioactive decay etc. [7]. The traditional statistical methods applied to survival analysis include parametric methods such as: the Kaplan-Meier (KM) estimator curve [8] and the Cox-proportional hazard (PH) models [9] while other methods apply non-parametric methods. Machine Learning (ML) does not rely on prior hypothesis unlike traditional explanatory statistical modeling techniques do [10]. Machine learning is a branch of artificial intelligence that allows computers to learn from past examples of data records ([11]; [12]). Machine learning techniques can be broadly classified into: supervised and Unsupervised techniques; the earlier involves matching a set of input records to one out of two or more target classes while the latter is used to create clusters or attribute relationships from raw, unlabeled or unclassified datasets [13].

Supervised machine learning algorithms can be used in the development of classification or regression models. Regression modeling is a supervised approach aimed at determining the numeric value of target class based on the values of a set of input records unlike classification which involves determining one out of many discrete values of a target class. This paper is focused at the application of regression modeling for the prediction of the survival of diabetes mellitus patients using supervised machine learning algorithms.

Diabetes Mellitus is a very serious disease affecting numerous Nigerians and has been identified as one of the major risk factors of many cardiovascular diseases. The limited number of experts available to manage patients compared to the number of cases attended to is gradually declining with a need for more effective means of performing decision-making during the treatment of diabetes mellitus. Related works in the healthcare information systems show that the increasing number of healthcare data requires the need of effective means of extracting information to aid the delivery of healthcare services to patients. There is a need for the development of a predictive model which will aid clinical decisions concerning continual treatment or alternative action affecting the survival of diabetes mellitus patients receiving treatment.

2. Related Works

Vijayalakshmi and Jenifer [14], applied data mining as a tool for analyzing clinical data records of both diabetic and non-diabetic patients. Data pertaining to 202 diabetic and 135 non-diabetic patients of the same attributes were collected from a nursing home research center in Trichy. WEKA software was used for applying various data mining techniques like Statistical analysis, Associative rule mining, and Clustering, Classification and subset evaluation. The results showed that among the algorithm used, the C4.5 decision tree was been found to be more accurate with an accuracy of 81%. The study focused on the classification of the survival of patients with DM.

Idowu et al. [15], applied machine learning to the prediction of the survival of pediatric HIV/AIDS patients. The machine learning algorithm used was the naïve Bayes' classifier. A 10-fold cross validation training technique was used to train the predictive model for survival classification of pediatrics HIV/AIDS patients using the data collected from south-western Nigeria hospital. The results of the study showed that the classifier was able to predict the survival of HIV/AIDS patients with an accuracy of 68%.

Idowu *et al*., [6], developed a predictive model for the survival of pediatric Sickle Cell Disease (SCD) using clinical variables. The predictive model was developed with fuzzy logic using three (3) clinical variables while the rules for the inference engine were elicited from expert pediatrician. The fuzzy logic based model was not validated using live clinical datasets. Moreover, relevant variables for SCD survival could have been easily identified using feature selection methods from a larger collection of variables monitored for pediatric SCD survival. The study did not simulate nor validate the rule-base of the classification model for SCD survival.

Sanakal and Jayakumari [16], developed a predictive model for the prognosis of diabetes using

the machine learning. The author used data collected from the University of California Illinois (UCI) repository consisting of 9 input attributes related to the clinical diagnosis of 768 patients. The author used the fuzzy c-means clustering and the support vector machines to formulate the predictive model for the diagnosis of diabetes. The results of the study showed that the fuzzy c-means clustering algorithm outperformed the SVM algorithm with an accuracy of 94.3% alongside a true positive rate of 95.4%.

Kumari and Chitra [17], developed a predictive model for the classification of diabetes disease using support vector machine (SVM). The author made use of the Pima Indian diabetes dataset which is a collection of medical diagnostic reports from 768 records of female patients who are at least 21 years old of Pima Indian heritage. The data contained 500 and 268 cases of patients that did not survived and those that survived respectively. A 10-fold cross validation technique was used to train the predictive model using the SVM classifier. The results of the study showed that the SVM had an accuracy of 78% with a true positive and true negative value of 80% and 77% respectively.

Agrawal *et al.,* [18], developed a predictive model for the classification of the survival of lung cancer patients. Data for the study was collected from the Surveillance, Epidemiology and End Results (SEER) Program with patients' data for survival of 6 months, 9 months, 1 year, 2 year and 5 years and consisting of 13 input variables. Different decision trees algorithms were used for the formulation of the predictive model, such as: C4.5 decision trees, random forest, Decision Stump and alternating decision trees. The decision trees algorithms used had the best accuracy with values of 73.61%, 74.45%, 76.80%, 85.45% and 91.35% for the 6 months, 9 months, 1 year, 2 year and 5 years survival dataset respectively.

3. Materials and Methods

The approach adopted in this study comprises of numbers of methods which include the identification of the required variables predictive of survival of diabetes mellitus; data collection method used in gathering the required data needed for model development; formulation of predictive models using the supervised machine learning algorithms proposed; the simulation of the predictive models using the WEKA simulation environment; and the performance evaluation metrics applied during model validation for the predictive models.

3.1 Data Identification and Collection

Following the review of related works of literature in the body of knowledge of survival of diabetes mellitus and the variables related to determine survival of diabetes mellitus, a number of variables were identified. The identified variables for determining survival of diabetes mellitus were validated by a physician interviewed with more than 10 years' experience in medicine before the data was collected. For analysis purpose data was collected from hospital case files of 29 patients undergoing treatment at a hospital located in the south-western part of Nigeria following the processing of health records' ethical clearance.

The information collected from the hospital was stored in a spreadsheet application – Microsoft Excel of the Microsoft Office 2013. Information collected consisted of the explanatory variables for the survival of diabetes mellitus as proposed by the expert for each patient. A description of the attributes contained in the dataset is presented in Table 1. Information about the aforementioned variables in Table 1 was collected from the information stored in the patient's case files and stored in an electronic format using the supervised machine learning algorithms. Some of the variables assessed were nominal while others assessed were numeric values such as the survival time which was assessed as the number of years for which the patient was on treatment.

3.2 Data Preprocessing

Following the collection of data of the 29 patients alongside the attributes consisting of 21 risk factors alongside the survival of diabetes mellitus. The data collected was checked for the presence of error in data entry including misspellings and missing data. The data was transformed into the attribute file format (.arff) for the purpose of the development of the predictive model for the survival of diabetes mellitus using the simulation environment. Figure 1 shows a screenshot of the format of the .arff used for model development in the Waikato Environment for Knowledge Analysis (WEKA) – a light-weight java application which composed of a suite of supervised and unsupervised machine learning tools.

Table 1: Identified variables for Survival of Diabetes mellitus

S/No.	Variable Names	Labels
1.	Gender	Male, Female
2.	Present Age (in years)	Numeric
3.	Highest Education	Primary, Polytechnic, Secondary, University, Nil
4.	Occupation	Driver, Trader, Banker, Student, Teacher, Retired, Nil, Cleaner
5.	Marital Status	Single, Married, Divorced
6.	Ethnicity	Yoruba, Hausa, Ibo
7.	Religion	Christian, Islam, Traditional, Nil
8.	Weight (in Kg)	Numeric
9.	Height (in meters)	Numeric
10.	Body Mass Index (BMI)	Numeric
11.	BMI Class	Underweight, Normal, Overweight and Obese
12.	Age at Diagnosis (in years)	Numeric
13.	Glucose Intake Level	Numeric
14.	Medicine Resistance	Very low, Low, Moderate, High
15.	Deflated EEB Level	Numeric
16.	Treatment	1–Diabohills, 2-Madhumehari, 3-Dbt SP, 4-Divoherb, 5-Magnetic Diabetes Belt, 6-Sanjecvani and 7-BGR – 34
17.	Medicine Effect	Increase, Decrease, None
18.	Body Chemistry	Slow, Moderate, Fast
19.	SBP (on drugs and after Treatment)	Numeric
20.	Change in SBP (in mmHg)	Numeric
21.	DBP (on drugs and after Treatment)	Numeric
22.	Change in DBP (in mmHg)	Numeric
21.	Treatment Time (in weeks)	Numeric
22.	Survival Time (in years)	Numeric

```
@relation diabetesTrainingData

@attribute Gender {male,female}
@attribute PresentAge numeric
@attribute HighestEducation {primary,polytechnic,secondary,university,nil}
@attribute Occupation {driver,trader,banker,student,teacher,retired,nil,cleaner}
@attribute MaritalStatus {single,married,divorced}
@attribute Ethnicity {yoruba,hausa,ibo}
@attribute Religion {christian,islam,traditional,nil}
@attribute Weight numeric
@attribute Height numeric
@attribute BMI numeric
@attribute BMI-class {underweight,normal,overweight,obese}
@attribute AgeDiagnosis numeric
@attribute GlucoseIntakeLevel numeric
@attribute MedicineResistance {very-low,low,moderate,high}
@attribute DeflatesEEBLevel numeric
@attribute Treatment1 {yes,no}
@attribute Treatment2 {yes,no}
@attribute Treatment3 {yes,no}
@attribute Treatment4 {yes,no}
@attribute Treatment5 {yes,no}
@attribute Treatment6 {yes,no}
@attribute Treatment7 {yes,no}
@attribute MedicineEffect {increase,decrease,none}
@attribute BodyChemistry {slow,moderate,fast}
@attribute SBP-OnDrugs numeric
@attribute SBP-AfterTreatment numeric
@attribute SBP-Change {increase,decrease,none}
@attribute DBP-OnDrugs numeric
@attribute DBP-AfterTreatment numeric
@attribute DBP-Change {increase,decrease,none}
@attribute TreatmentTime numeric
@attribute Survival-time numeric

@data
female,56,polytechnic,driver,divorced,hausa,christian,71,1.7,24.56747405,normal,55,158,moderate,86,yes,yes,yes,no,no,no,no,decrease,moderate,140,120,decrease,70,80,increase,8,1
female,61,nil,trader,divorced,yoruba,islam,63,1.77,20.10916403,normal,60,150,low,111,yes,no,yes,no,yes,no,no,decrease,fast,140,120,decrease,90,80,decrease,24,1
male,56,secondary,trader,married,hausa,islam,70,1.6,27.34375,overweight,55,158,low,86,yes,yes,yes,no,no,no,no,decrease,moderate,140,110,decrease,70,70,none,8,1
female,62,nil,trader,divorced,yoruba,islam,72,1.6,28.125,overweight,60,150,moderate,112,yes,no,yes,no,yes,no,no,decrease,moderate,140,120,decrease,90,80,decrease,20,2
female,61,primary,trader,married,yoruba,?,75,1.7,25.95155709,overweight,59,403,high,120,yes,no,yes,no,yes,no,no,decrease,fast,160,120,decrease,90,80,decrease,4,2
female,31,university,banker,married,yoruba,christian,69.1,1.72,23.35722012,normal,28,360,high,110,yes,no,yes,yes,yes,no,no,decrease,fast,160,110,decrease,80,70,decrease,12,3
```

Normal text file | length : 6,210 lines : 66 | Ln : 18 Col : 31 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

Figure 1: Arff File containing Identified Attributes

The dataset collected for the purpose of the development of the predictive model for the survival of diabetes mellitus was stored in .arff in the name diabetes Training Data.arff while the numbers of attributes listed in the attribute section were 22 including the target attribute. Following this, the values of the risk factors for the record of the 29 patients considered for this study was provided. The arff file is composed of three parts, namely:

a. The relation name section which contains the tag @relation *diabetes Training Data*, used to identify the name of the relation (or file) that contains the data needed for simulation. This section is located at the first line of the file and the tag 'name' following @relation must always be the same as the file name else the file loader of the simulation environment will cease to open the file. This section is followed in the next line by the attribute names section;
b. The attribute names section which contains the tag @attribute, *attribute name label* was used to identify the attributes that describe the dataset stored in the .arff file needed for simulation. Each attribute name alongside its labels is stated following the @relation tag on each line. The label can be a set of values inserted between brackets or a descriptor (e.g. date, numeric etc.). The last attribute is identified as the target class (survival of diabetes mellitus) while the previous attributes are the variables for the survival of diabetes mellitus. This section is followed in the next line by the data section; and
c. The data section which contains the tag @data followed in the next line by the values of the attributes for each record of the survival of diabetes mellitus separated by a comma. Each value was listed on a row for each record in the same order as the attributes were listed in the attribute names section. The values inserted into each record must be the same values defined in each respective attribute; if there is an error in spelling or a label not defined is inserted then the file loader of the simulation environment will fail to load the file.

3.3 Model Formulation

Supervised machine learning algorithms are Black-boxed models, thus it was not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – survival of diabetes mellitus). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development. Systems that construct regression models are one of the commonly used tools in data mining. Such systems takes as input a collection of cases, each belonging to a numeric value for the target class and described by its values for a fixed set of attributes, and output a regression model that can accurately predict the value of the survival time. Supervised machine learning algorithms make it possible to assign a set of records (diabetes mellitus survival indicators) to a target classes – the survival time of diabetes mellitus.

For any supervised machine learning algorithm proposed for the formulation of a predictive model, a mapping function was used to express the general expression for the formulation of the predictive model for the classification of survival time of diabetes mellitus using a non-deterministic expression. The historical dataset S which consists of the records of patients containing attributes representing the set of identified factors (i number of input variables for j patients), X_{ij} alongside the respective target variable (survival time of diabetes mellitus) represented by the variable Y_j – the survival time of diabetes mellitus for the jth individual in the j records of data collected from the hospital selected for the study. Equation 1 shows the mapping function that describes the relationship between the classification factors and the target class – survival time of diabetes mellitus patients

$$\varphi: X \rightarrow Y \quad \rightarrow \quad defined\ as: \varphi(X) = Y \tag{1}$$

The equation shows the relationship between the set of factors represented by a vector, X consisting of the values of i variables and the label Y which defines the survival time of diabetes mellitus for each patient as expressed in equation 2. Assuming the values of the set of variable for a patient is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each variable, i = 1 to i; then the mapping φ used to represent the predictive model for patient performance maps the variables of each individual to their respective survival of diabetes mellitus according to equation 2.

$$\varphi(X) = \mathbb{N} \quad where \in \mathbb{R}\ (real\ number) \tag{2}$$

The developed predictive model for the survival time of diabetes mellitus was formulated using the support vector machine (SVM) and the multi-layer perceptron (MLP) algorithms. The algorithms

considered in this study fall under the class of perceptron network systems since input values are fired into nodes with synaptic weights assigned – inputs are sum of products of weights w_i and input x_i, equation (3) shows the expression.

$$\sum_{k=1}^{i} w_k x_k = w_1 x_1 + w_2 x_2 + \cdots + w_i x_i =< w.x \qquad (3)$$

3.4 Model Simulation Techniques

Following the identification of the supervised machine learning algorithms that was needed for the formulation of the predictive model for the survival of diabetes mellitus, the simulation of the predictive model was performed using the data collected which consisted of patients records containing information about the input variables and their respective value of survival time of diabetes mellitus collected from the hospital located in south-western Nigeria. The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model. The dataset collected was divided into two parts: training and testing data – the training data was used to formulate the model while the test data was used to validate the model.

The process of training and testing predictive model according to literature is a very difficult experience especially with the various available validation procedures. For this problem, it was natural to measure the model's performance in terms of the error rate. The error rate being the proportion of errors made over a whole set of instances, and thus measured the overall performance of the classifier. The error rate on the training data set was not likely to be a good indicator of future performance; because the models were been learned from the very same training data.

For this study the cross-validation procedure was employed, which involved dividing the whole datasets into a number of folds (or partitions) of the data. Each partition was selected for testing with the remaining k – 1 partitions used for training; the next partition was used for testing with the remaining k – 1 partitions (including the first partition used or testing) used for training until all k partitions had been selected for testing. The error rate recorded from each process was added up with the mean the mean error rate recorded. The process used in this study was the stratified 10-fold cross validation method which involves splitting the whole dataset into ten partitions.

3.5 Model Validation for Performance Evaluation

The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. The mean absolute error (MAE) is another useful measure widely used in model evaluations. While they have both been used to assess model performance for many years, there is no consensus on the most appropriate metric for model errors. In the field of geosciences, many present the RMSE as a standard metric for model errors.

While the MAE gives the same weight to all errors, the RMSE penalizes variance as it gives errors with larger absolute values more weight than errors with smaller absolute values. When both metrics are calculated, the RMSE is by definition never smaller than the MAE. Assuming there are n samples of model errors ϵ calculated as $(\epsilon_i,\ i = 1, 2, \ldots, n)$. The uncertainties brought in by observation errors or the method used to compare model and observations are not considered here. We also assume the error sample set ϵ is unbiased. The RMSE and the MAE are calculated for the data set as:

$$MAE = \frac{1}{2}\sum_{i=1}^{n} |\epsilon_i| \qquad (4)$$

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n} \epsilon^2{}_i} \qquad (5)$$

The underlying assumption when presenting the RMSE is that the errors are unbiased and follow a normal distribution. Thus, using the RMSE or the standard error (SE) helps to provide a complete picture of the error distribution. Condensing a set of error values into a single number, either the RMSE or the MAE, removes a lot of information. The best statistics metrics should provide not only a performance measure but also a representation of the error distribution. The MAE is suitable to describe uniformly distributed errors. Because model errors are likely to have a normal distribution rather than a uniform distribution, the RMSE is a better metric to present than the MAE for such a type of data.

Correlation coefficient is a measure of association between two variables, and it ranges between –1 and 1. If the two variables are in perfect linear relationship, the correlation coefficient will be either 1 or –1. The sign depends on whether the variables are positively or negatively related. The correlation coefficient is 0 if there is no linear relationship between the variables. Two different types of correlation coefficients are in use. One is called the Pearson product moment correlation coefficient, and the other is called the Spearman rank correlation coefficient, which is based on the rank relationship between variables. Given paired measurements(X_1, Y_1),(X_2, Y_2),(X_3, Y_3), ..., (X_n, Y_n), the Pearson product moment correlation coefficient is a measure of association given by:

$$r_p = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (6)$$

where: $\bar{X}$ *and* $\bar{Y}$ are the sample means of $X_1, X_2. X_3 \ldots\ldots X_n$ and $Y_1, Y_2. Y_3 \ldots\ldots Y_n$, respectively.

4. Results

This section presents the results of the methods that were applied for the development of the predictive model for the survival of diabetes mellitus. The results presented were that of the data collection, model formulation and simulation results using the WEKA software following the results of the model validation of the predictive model for survival of diabetes mellitus. A thorough investigation into the analysis of the description of the dataset collected was initially performed in order to understand the distribution of the values of the variable for each survival of diabetes mellitus among the patients selected for this study using the minimum and maximum values, and the mean and standard deviation of the data distribution. The numeric variables identified and collected for the study were also discretized into nominal values so as to reduce the computational complexity associated with numeric variable.

4.1 Data Description

Table 2 shows the description of the nominal variables while Table 3 shows the distribution of the numeric variables. From the description shown in Table 2, there were more female than male respondents owing to a percentage of 65.5% and 34.5% of patients for female and male respectively. The results of the education qualification showed that majority had secondary education (24.1%) followed by those with polytechnic education (20.7%) and primary education (17.2%). The results further showed that majority of the patients were married with a proportion of 65.6% followed by divorced patients with a proportion of 31.1% while the results of the ethnicity showed that majority of the patients were Yoruba with a proportion of 51.7% followed by the Ibo and Hausa with a proportion of 27.6% and 20.7% respectively. The results of the study also showed that majority of the patients were Christians and Muslims with proportion of 41,1% each while the results of the body mass index (BMI) showed that majority of the patients were overweight with a proportion of 48.2% followed by those that were normal and obese with proportion of 34.5% and 13.8% respectively.

The results further revealed that information regarding the variables used to monitor the survival of diabetes mellitus patients highlights that the majority of the patients had moderate and high resistance to the treatment administered with proportion of 34.5% each followed by those with low and very low resistance with proportion of 17.2% and 13.8% respectively. The treatment result showed that majority of the patients were given treatment 3 (Dbt SP) with a proportion of 79.3% followed by those administered treatment 2 (Madhumehari) with a proportion of 72.2%, followed by those administered treatment 5 (Magnetic diabetes belt) with a proportion of 65.5% and treatments 1 (Diabohills) and 4 (Divoherb) with proportions of 44.8% and 37.9% respectively. The results further established that the majority of the patients had a decrease in systolic blood pressure (SBP) after treatment compared to when on drugs with a proportion of 70% while 10.3% had an increase while majority of the patients had decrease in diastolic blood pressure (DBP) after treatment compared with when on drugs with a proportion of 58.6% while 10.3% had an increase in DBP.

From the description shown in Table 3, the analysis of the numeric datasets is presented showing the values of the minimum, maximum, mean and standard deviation of each variable presented in the dataset. The results of showed that the minimum and maximum ages of patients were 11 and 69 years while the minimum and maximum age at diagnosis were 7 and 61 years with average ages of 58 and 49 years for their present age and age at diagnosis.

Table 2: Description of the nominal variables in the dataset

Variables	Labels	Frequency	Percentage (%)
Gender	Male	10	34.5
	Female	19	65.5
Highest Education	Primary	5	17.2
	Secondary	7	24.1
	Polytechnic	6	20.7
	University	4	13.8
	None	7	24.1
Marital Status	Single	1	3.4
	Married	19	65.5
	Divorced	9	31.1
Ethnicity	Yoruba	15	51.7
	Hausa	6	20.7
	Ibo	8	27.6
Religion	Christian	12	41.4
	Islam	12	41.4
	Traditional	1	3.4
	None	1	3.4
	Missing	3	10.4
BMI Class	Underweight	1	3.4
	Normal	10	34.5
	Overweight	14	48.2
	Obese	4	13.8
Medicine Resistance	Very low	4	13.8
	Low	5	17.2
	Moderate	10	34.5
	High	10	34.5
Treatment	Treatment 1	13	44.8
	Treatment 2	21	72.2
	Treatment 3	23	79.3
	Treatment 4	11	37.9
	Treatment 5	19	65.5
	Treatment 6	2	6.9
	Treatment 7	3	10.3
Medicine Effect	Increase	0	0.0
	Decrease	28	96.6
	None	1	3.4
Body Chemistry	Slow	1	3.4
	Moderate	13	44.8
	High	15	51.8
SBP Change	Increase	3	10.3
	Decrease	20	70.0
	None	4	13.8
	Missing	2	6.9
DBP Change	Increase	3	10.3
	Decrease	17	58.6
	None	6	20.7
	Missing	3	10.3

Table 3: Description of the numeric variables in the dataset

Variables	Minimum	Maximum	Mean	Standard Deviation
Present Age (in years)	11	69	58.21	13.276
Weight (in Kg)	30	92	69.92	10.841
Height (in meters)	1.3	1.8	1.64	0.097
BMI	18	40	26.19	4.493
Age at Diagnosis (in years)	7	61	49.10	12.310
Glucose Intake Level	150	417	251.21	95.164
Deflated EEB Level	76	264	126.17	47.07
SBP (on drugs in mmHg)	100	180	144.36	21.596
SBP (after medication in mmHg)	110	140	124.44	9.740
DBP (on drugs in mmHg)	60	120	88.52	15.62
DBP (after medication in mmHg)	60	90	78.15	6.225
Treatment Time (in weeks)	1.5	364	36.24	71.564
Survival Time (in years)	1	22	9.10	5.802

The results also showed that the minimum and maximum weights were 30 and 93 kg while the minimum and maximum heights were 1.3 and 1.8 metres respectively. In addition the results revealed that the minimum and maximum SBP were 100 and 180 when on drugs and 110 and 140 after treatment while the minimum and maximum DBP were 60 and 120 when on drugs and 60 and 90 after treatment. The results also observed that the minimum and maximum survival times were 1 and 22 years with an average survival time of 22 years. Figure 2 shows a plot of the distribution of the survival time of the patients from the lowest to the highest survival time (in years) based on the results of the study.

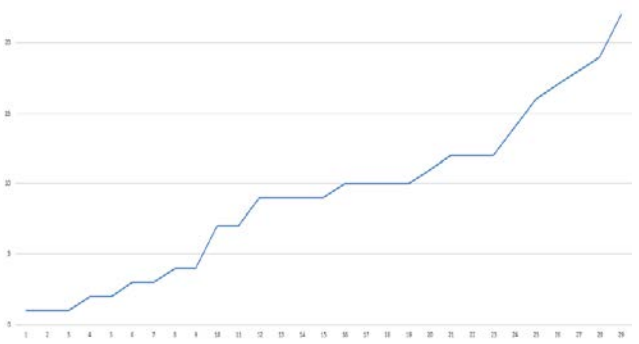

Figure 2: Graphical plot of survival time (in years) for all patients

Following this, the results of the model formulation and simulation process for the development of the predictive model for the survival time of diabetes mellitus were presented.

The performance of the predictive models for patient performance developed using the machine learning algorithms were evaluated in order to identify the most effective and efficient predictive model for the survival of diabetes mellitus.

4.1 Results of Model Simulation and Validation

Two different supervised machine learning algorithms were used to formulate the predictive model for the survival of diabetes mellitus, namely: support vector machines and the multilayer perceptron algorithms. They were used to train the development of the prediction model using the dataset containing 29 patients' records. The simulation of the prediction models was done using the Waikato Environment for Knowledge Analysis (WEKA). The multilayer perceptron was implemented using the *Multilayer Perceptron* algorithm while the support vector machine algorithm was implemented using the *SMOreg* algorithm both made available in the functions

classifier on the WEKA Explorer environment. The models were trained using the 10- fold cross validation method which splits the dataset into 10 subsets of data – while 9 parts are used for training the remaining one is used for testing; this process is repeated until the remaining 9 parts take their turn for testing the model.

Following the simulation of the predictive model for the survival time of diabetes mellitus using the support vector machines, the evaluation of the performance of the model following validation using the 10-fold cross validation method was recorded. Figure 3 shows the screenshot of the results of the predictions made by the support vector machine algorithm for the 29 instances of data collected from the patients considered for the study. The figures show the correct and incorrect classifications made by the algorithm.

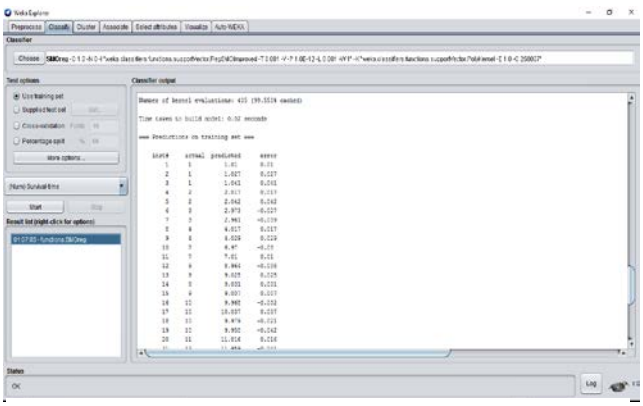

Figure 3: Graphical plot of the actual and predicted values of the SVM model

Figure 4(a) shows the graphical plot of the actual and predicted values of the support vector machines while figure 4(b) shows a graphical plot of the error values of each prediction made by the support vector machine algorithm. The results of the study revealed that the minimum error rate recorded was -0.042 while the maximum error rate was 0.042 with a mean square error (MSE) value of 0.000827.

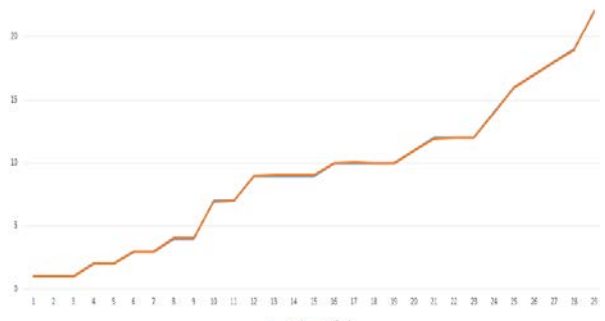

Figure 4(a): Graphical plots of the actual and predicted values SVM

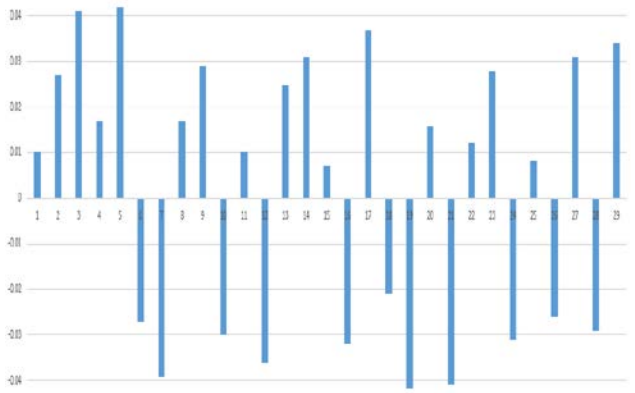

Figure 4(b): Graphical plots of the errors value of each prediction made by SVM algorithm

The performance of the predictive models for patient performance developed using the machine learning algorithms were evaluated in order to identify the most effective and efficient predictive model for the survival of diabetes mellitus.

Following the simulation of the predictive model for the survival of diabetes mellitus using the multilayer perceptron, the evaluation of the performance of the model following validation using the 10-fold cross validation method was recorded.

Figure 5 shows the screenshot of the results of the predictions made by the multilayer perceptron algorithm for the 29 instances of data collected from the patients considered for the study. The figures show the correct and incorrect classifications made by the algorithm. Figure 6 shows the graphical plot of the actual and predicted values of the multilayer perceptron for each prediction made by the multilayer perceptron algorithm.

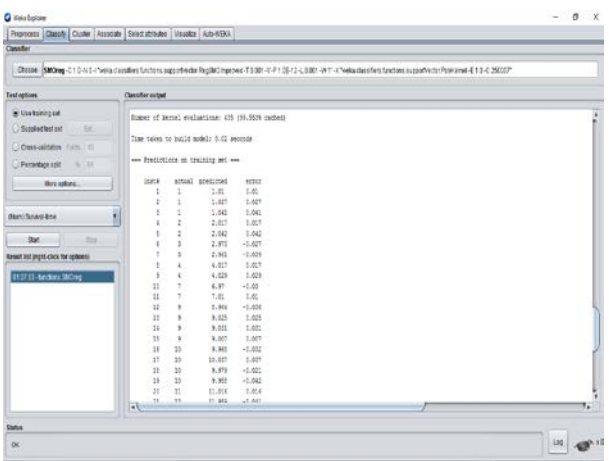

Figure 5: Screenshot of support vector machines results on dataset

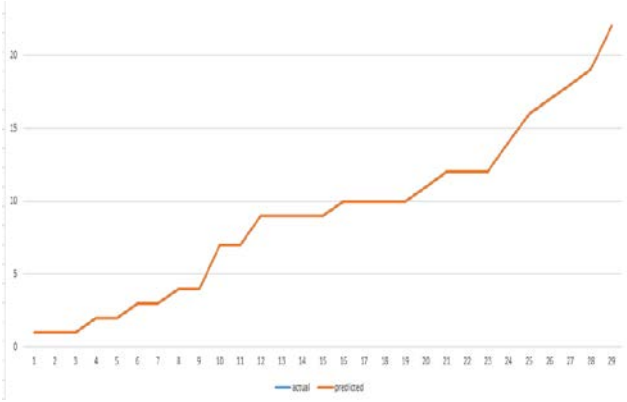

Figure 6: Graphical plot of the actual and predicted values of the MLP model

The results of the study revealed that the MLP predicted the survival time of the diabetes mellitus patients without error.

4.2 Discussion of Results

The result of the performance evaluation of the SVM and MLP algorithms are presented in Table 4. The results showed that predictive model developed by the support vector machine algorithm for the survival of diabetes mellitus was completed within 0.02 seconds while the model developed by the multilayer perceptron was developed within 0.77 seconds owing to the fact that the model complexity of the MLP resulted in a longer time of processing the predictive model for the survival of diabetes mellitus patients.

The results also revealed that the support vector machine showed a mean absolute error value of 0.0267 compared with the value of 0 performed by the multilayer perception.

The results further showed that the root mean square error value of the support vector machine has a value of 0.0287 compared to the value of 0 for the multilayer perceptron while the relative absolute error was 0.5915 for the support vector machine compared to the value of 0.0009 for the multilayer perceptron. It was revealed that both perceptron-based networks were able to predict the survival time of diabetes mellitus patients whom were receiving treatment.

The performance of the multilayer perceptron and the support vector machine was very close; the support vector machine developed the predictive

Table 4: Summary of the results of performance evaluation for the machine learning algorithms selected

Machine Learning Algorithm Used	PERFORMANCE EVALUATION METRICS				
	Model build Time (s)	Correlation Coefficient	Mean absolute error (MAE)	Root mean square error (MSE)	Relative absolute error (%)
Support Vector Machines	0.02	1.00	0.0267	0.0287	0.5915
Multi-Layer Perceptron	0.77	1.00	0.0000	0.0000	0.0009

model for the survival of diabetes mellitus patients in a shorter time compared to that of the multilayer perceptron algorithm. The error rates recorded by the support vector machine were significant while that of the multilayer perceptron were very insignificant. Therefore, the multilayer perceptron algorithm performed better than the support vector machine algorithm in terms of error rates while the support vector machine algorithm had a better processing speed.

5. Conclusion

This study focused on the development of a prediction model using identified variable predictive of survival of diabetes mellitus in order to classify the survival of diabetes mellitus in selected patients for this study. The identified variables for determining survival time of diabetes mellitus were validated by a physician interviewed with more than 10 years' experience in medicine, following which the dataset on the distribution of the variable determinant of survivability of diabetes mellitus among 29 patients were collected from patient case files in a hospital located in south-western part of Nigeria.

The dataset containing information about the variables collected from the patients file was used to formulate predictive models for the survival of diabetes mellitus patients using Support Vector Machine and Multi-layer Perception algorithms. The predictive model development using the algorithms

were formulated and simulated using the WEKA software.

The development of a predictive model for predicting the survival of diabetes mellitus given the values of variables was developed using dataset collected. Thirty two (21) variables were identified by the medical expert to be necessary in predicting diabetes mellitus in patients for which a dataset containing information of 29 patients alongside their respective diabetes mellitus survival time provided. A 10-fold cross validation method was used to train the predictive model developed using the machine learning algorithms and the performance of the models evaluated.

The results of the study revealed that the multi-layer perceptron algorithm proved to be better in terms of error rates while the support vector machine had a better processing time for model development for predicting the survival of diabetes mellitus in Nigerian patients. Following the development of the predictive model for the survival time of diabetes mellitus patients receiving treatment, the multilayer perceptron algorithm was proposed due to the understanding of the relationship between the attributes and the survival time. The model can also be integrated into existing Health Information System (HIS) which captures and manages clinical information which can be fed to the diabetes mellitus predictive model thus improving the decisions affecting the patient's outcome and the real-time assessment of clinical information affecting the survival time of diabetes mellitus patients.

References

[1] Chijioke A., Adamu, A.N. and Makusidi, A.M. (2010). Mortality Patterns among Type 2 Diabetes Mellitus Patients in Ilorin, Nigeria. *JEMDSA 15*: 79–82.

[2] World Health Organization (WHO) (2016). World Diabetes Report 2016. Available from http://www.who.int/news-room/ fact-sheets/details/diabetes. Accessed 25th March 2018.

[3] Aguocha, B.U., Ukpabi, J.O. and Onyeonoro, U.U. (2013). Pattern of Diabetic Mortality in a Tertiary Health Facility in South-Eastern Nigeria. *African Journal of Diabetes Medicine 21*: 14–16.

[4] International Diabetes Federation Editorial Team. Mortality. In: Guariguata L, Nolan T, Beagley J, Linnenkamp U, Jacqmain O (Eds). Diabetes Atlas, 6th edition. Brussels: International Diabetes Federation (IDF), 2003: 49. www.idf.org/diabetesatlas.

[5] Ojobi, J.E., Odoh, G., Aniekwensi, E. and Dunga, J. (2016). Mortality among Type 2 Diabetic In Patients in a Nigerian Tertiary Hospitals. *African Journal of Diabetes Medicine*, Vol. 24(2). Pp. 17-20.

[6] Idowu, P.A., Aladekomo, T.A., Williams, K.O. and Balogun, J.A. (2015). Predictive Model for Likelihood of Sickle Cell Aneamia (SCA) among Pediatric Patients using Fuzzy Logic. *Transactions in networks and communications 31*(1): 31 – 44.

[7] Dimitoglou, G. and Adams, J. A. (2012): Comparison of the C4.5 and the Naïve bayes classifier for the prediction of lung cancer survivability. Journal of Computing 4(8): 1 -13

[8] Kaplan, E.L., and Meier, P. (1958). Non Parametric Estimation from Incomplete Observations. *Journal of American Statiatical Association.* Vol.53 (282) Pp. 457-481

[9] Cox, D.R. (1972). Regression Models and Life-Tables. *Royal Statistical Society*, Series B, Vol. 34(2) pp. 187-220.

[10] Waijee, A., Mukherjee, A. and Singal, A. (2013b). Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine. *BMJ Open 3*(8): 1 – 7.

[11] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning 1*: 81-106.

[12] Cruz, J.A. and Wishart, D.S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics 2*: 59 – 75

[13] Michell, T. M. (1997). Machine Learning. Mc Graw.Hill, Inc. New York, NY, USA.

[14] Vijayalakshmi, N. and Jennifer T. (2017). An Analysis of Risk Factors for Diabetes using Data Mining Approach. *International Journal of Computer Science and Mobile Computing*, Vol.6 (7) pp.166-172

[15] Idowu, P.A., Agbelusi, O. and Aladekomo, T.A. (2016). The Prediction of Pediatric HIV/AIDS Patients' Survival: A Data Mining Approach. *Asian Journal of Computer and Information Systems 4*(3): 87 – 94.

[16] Sanakal, R. and Jayakumari, T. (2014). Prognosis of Diabetes using Data Mining Approach – Fuzzy C Means Clustering and Support Vector Machine. *International*

Journal of Computer Trends and Technology (IJCTI) 11(2): 94 – 98.

[17] Kumari, V.A. and Chitra, R. (2013). Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Application 3*(2): 1897 – 1801.

[18] Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L. and Choudhary, A. (2012). Lung Cancer Survival Prediction Using Ensemble Data Mining on SEER Data. *Journal of Scientific Programming 20*: 29 – 42.