Computational Biology and Bioinformatics
2020; 8(2): 52-61
http://www.sciencepublishinggroup.com/j/cbb
doi: 10.11648/j.cbb.20200802.14
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Survival Model for Diabetes Mellitus Patients' Using Support Vector Machine

Samson Alobalorun Bamidele[1], Adanze Asinobi[2], Ngozi Chidozie Egejuru[3], Peter Adebayo Idowu[4]

[1]Department of Computer, Library and Information Science, Kwara State University, Malete, Nigeria

[2]Department of Paediatrics College of Medicine, University of Ibadan, Ibadan, Nigeria

[3]Department of Computer Science, Hallmark University, Ijebu Itele, Nigeria

[4]Department of Computer Science & Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

Email address:
paidowu1@yahoo.com (P. A. Idowu)

To cite this article:
Samson Alobalorun Bamidele, Adanze Asinobi, Ngozi Chidozie Egejuru, Peter Adebayo Idowu. Survival Model for Diabetes Mellitus Patients' Using Support Vector Machine. *Computational Biology and Bioinformatics*. Vol. 8, No. 2, 2020, pp. 52-61.
doi: 10.11648/j.cbb.20200802.14

Received: May 30, 2020; **Accepted**: October 23, 2020; **Published**: November 4, 2020

Abstract: This study developed a model for the survival of diabetes mellitus patients in Nigeria. The study identified the variables monitored during the treatment of diabetes mellitus patients, formulated, and validated the predictive model for the survival time of diabetes mellitus patients. In order to achieve the aim of this study, structured interview with professional physicians so as to identify the variables for the survival time of diabetes mellitus with historical datasets were collected based on the variables monitored during treatment. The model was formulated using the support vector machine based on the variables identified and simulated using the WEKA Software using the historical datasets for training the model. The results showed that data collected from 29 patients at a hospital located in south-western Nigeria consisting of 32 attributes with a target class containing information about the survival time of each diabetes mellitus patient. The study concluded that the model can also be integrated into existing Health Information System (HIS) which captures and manages clinical information which can be fed to the predictive model thus improving the decisions affecting the patient's outcome and the real-time assessment of clinical information affecting the patient's survival of diabetes.

Keywords: Support Vector Machine, Diabetes Mellitus, Survival, Model, Predictive

1. Introduction

Diabetes mellitus now constitutes the highest morbidity and mortality of all chronic non-communicable diseases (NCDs) in Africa. In Nigeria, diabetes accounts for 3–15% of medical admissions in most health facilities [2, 11]. People living with type 2 diabetes are more vulnerable to various forms of both short- and long-term complications, which often lead to their premature death [3]. According to a report by International Diabetes Federation [8], close to half (48%) of deaths due to diabetes are in people under the age of 60 years. Approximately 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 [15]. 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.

Survival Analysis deals with the application of methods to estimate the likelihood of an event (death, survival, decay, child-birth etc.) occurring over a variable time period (Dimitologlou *et al.,* 2012); in short, it is concerned with studying the time between entry to a study and a subsequent event (such as death). The traditional statistical methods applied in the area of survival analysis include the Kaplan-Meier (KM) estimator curve (Kaplan *et al.,* 1958) and the Cox-proportional hazard (PH) models [4]. These methods apply parametric methods in estimating survival parameters for a group of individuals. Other methods applied in traditional statistical methods also include the use of non-parametric models. The Kaplan-Meier method allows for an estimation of the proportion of the population of people who survive a given length of time under some circumstances.

The cox model is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables.

Machine learning is a branch of artificial intelligence that allows computers to learn from past examples of data records [5, 12]. Machine learning does not rely on prior hypothesis unlike traditional explanatory statistical modeling techniques do [14]. Machine learning has found great importance in the area of predictive modeling in medical research especially in the area of risk assessment, risk survival and risk recurrence. Machine learning techniques can be broadly classified into: supervised and unsupervised learning 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 (Mitchell, 1997). 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 and this is the focus of this paper.

2. Related Works

Li *et al.* (2010) developed a predictive model for renal graft status and survival period using the Byes' Net Classifier. Data was collected from the University of Toledo Medical Center Hospital patients as reported to the United Network Organ Sharing, and had 1228 patient records for the period covering 1987 through 2009. The Bayes net classifiers were developed using the Weka machine learning software workbench. Two separate classifiers were induced from the data set, one to predict the status of the graft as either failed or living, and a second classifier to predict the graft survival period. The classifier for graft status prediction performed very well with a prediction accuracy of 97.8% and 68.2% and true positive values of 0.85 and 0.988 for the class representing those instances with kidneys failing during the first year following transplantation for the first and second classifiers respectively. The simulation results indicated that it is feasible to develop a successful Bayesian belief network classifier for prediction of graft status, but not the graft survival period, using the information in UNOS database.

Agrawal *et al.* [1], developed a predictive model for the classification of the survival of the survival of lung cancer patients. Data for the study was collected from the Surveillance, Epidemiology and End Results (SEER) Program containing patients' data for survival of 6 months, 9 months, 1 year, 2 year and 5 years 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 accuracies 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.

Kumari and Chitra [9], developed a predictive model for the classification of diabetes disease using support vector machine (SVM)s. The study made use of the Pima Indian diabetes dataset, donated by Vincent Sigillito which is a collection of medical diagnostic reports from 768 records of female patients at least 21 years old of Pima Indian heritage, a population living near Phoenix, Arizona, USA. The data contained 500 and 268 cases of patients that did not survived and those that survived respectively. The 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.

Sanakal and Jayakumari [13], developed a predictive model for the prognosis of diabetes using the fuzzy c-means clustering and the support vector machines. The study 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 study 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%.

Idowu *et al.* [7], developed a predictive model for the survival of pediatric sickle cell disease (SCD) using clinical variables. The predictive model was developed using a fuzzy logic based model using three (3) clinical variables. The model developed using the fuzzy logic model was not validated using live clinical datasets. Relevant variables for SCD survival could have been easily identified using feature selection methods.

Idowu *et al.* [6], applied supervised machine learning algorithm to the prediction of the survival of pediatric HIV/AIDS patients. The machine learning algorithms used was the naïve Bayes' classifier. The 10-fold cross validation training technique was used to train the predictive model for survival classification of pediatrics HIV/AIDS patients data collected from south-western Nigeria. The results of the study showed that the classifier was able to predict the survival of HIV/AIDS patients with an accuracy of 68%.

3. Methods

To develop the predictive model for the survival of diabetes mellitus in a well-detailed manner. The methodology consists of a sequence of methods/techniques which started with the identification of the variables predictive of survival of diabetes mellitus alongside the data collection method used in gathering the required data needed for model development. The historical data collected contained records of patients consisting of their respective values for each identified variables as inputs alongside the target variable (survival time of diabetes mellitus) as the output variable.

The machine learning algorithms used in formulating the

predictive model was proposed alongside the process of model development using the historical data for training and testing the predictive model for the survival of diabetes mellitus.

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 from the hospital located in the south-western part of Nigeria. Data were collected from 29 patients undergoing treatment at a hospital located in the south-western part of Nigeria from hospital case files following the processing of health records' ethical clearance. The information collected from the hospital was collected and stored in a spreadsheet application – Microsoft Excel of the Microsoft Office 2013. Information collected from the patients contained the explanatory variables for the survival of diabetes mellitus as proposed by the cardiologist for each patient. A description of the attributes contained in the dataset is presented in Table 1.

3.2. Data-Preprocessing

Following the collection of data from the 29 patients alongside the attributes (32 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 composed of a suite of supervised and unsupervised machine learning tools.

Table 1. *Identified variables for determining Diabetes mellitus.*

S/N	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 metres)	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

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 *diabetesTrainingData.arff* while the number of attributes listed in the attribute section were 33 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.

3.3. Formulation of Predictive Model for Diabetes Mellitus Patients' Survival

Systems that construct regression models take 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. Supervised machine learning algorithms are Black-boxed models, thus it is 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.

For any supervised machine learning algorithm proposed for the formulation of a predictive model, a mapping function can be used to easily express the general expression for the formulation of the predictive model for the classification of survival of diabetes mellitus – this is as a result that most machine learning algorithms are black-box models which use evaluators and not power series/polynomial equations. The historical dataset S which consists of the records of patients containing fields representing the set of classification factors (i number of input variables for j patients), X_{ij} alongside the respective target variable (survival 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 3.1 shows the mapping function that describes the relationship between the classification factors and the target class – survival time of diabetes mellitus patients.

$$\varphi: X \rightarrow Y \tag{1}$$

$$defined\ as: \varphi(X) = Y$$

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 3.2. 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 3.2.

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

The developed predictive model for the survival of diabetes mellitus was used to develop the predictive model for determining the survival time of diabetes mellitus directly just by training the model with the support vector machine algorithms.

3.4. Model Simulation Process and Environment

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 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.

In order to predict the performance of the model on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the model. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data. It was important that the test dataset was not used in any way to create the classifier since the machine learning classifiers involve two stages: one to come up with a basic structure of the predictive model and the second to optimize parameters involved in that structure.

i. 10-fold cross validation technique

The process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. The challenge here is the need to be able to find a good classifier by using as much of the whole historical data as possible for training; to obtain a good error estimate and use as much as possible for model testing. It is a common trend to holdout one-third of the whole historical dataset for testing and the remaining two-thirds for training.

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.

ii. Simulation environment

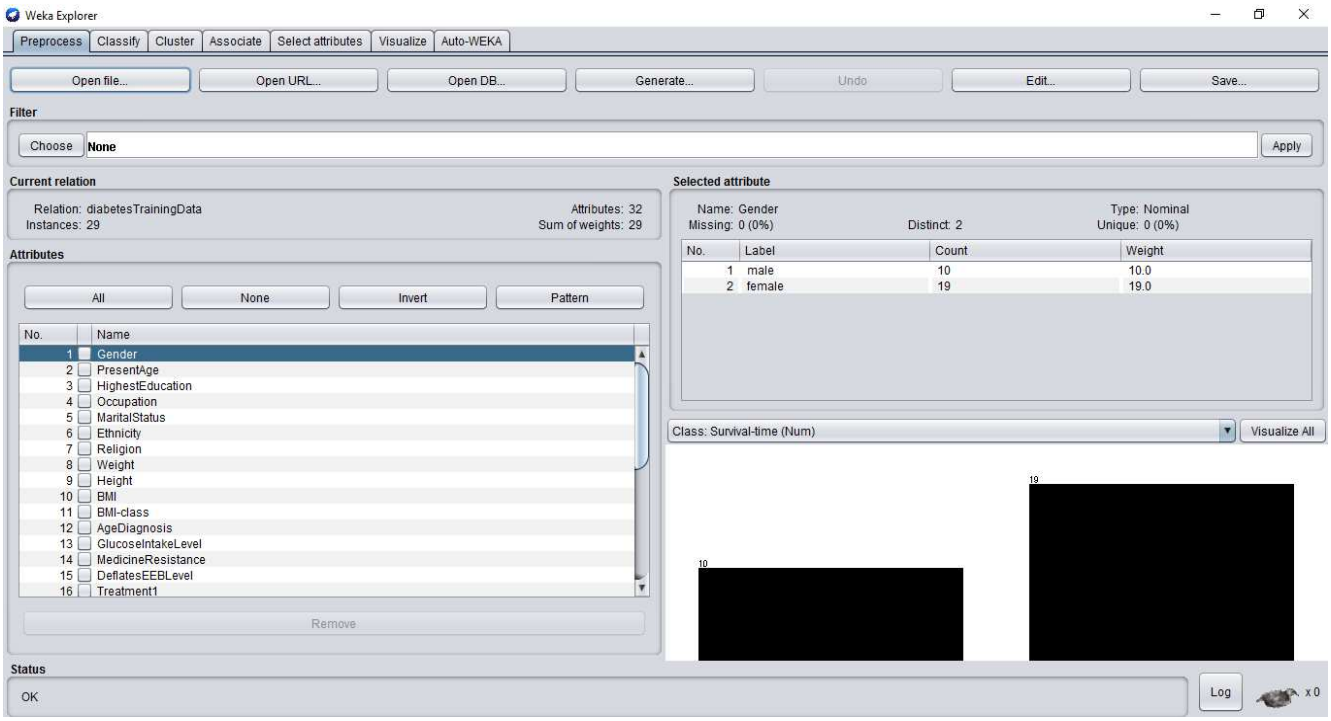

Figure 1. *The WEKA Explorer Interface Environment for Simulation.*

Weka is open source software under the GNU General Public License. The system was developed at the University of Waikato in New Zealand. Weka stands for the Waikato Environment for Knowledge Analysis. The software is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system was written using object-oriented language, Java. There are several different levels at which Weka can be used. Weka provides implantations of state-of-the-art data mining and machine learning algorithms. Weka contains modules for data preprocessing, classification, clustering and association rule extraction for market basket analysis (see Figure 1). The main features of Weka include:

a) 49 data preprocessing tools;
b) 76 classification/regression algorithms;
c) 8 clustering algorithms;
d) 15 attribute/subset evaluators + 10 search algorithms for feature selection;
e) 3 algorithms for finding association rules; and
f) 3 graphical user interfaces, namely:
i. The Explorer for exploratory data analysis;
ii. The Experimenter for experimental environment; and
iii. The Knowledge Flow, a new process model inspired interface.

Before subjecting the historical datasets containing the values of the variables alongside the survival of diabetes mellitus for each patient's record in the original dataset; there was the need of storing the dataset according to the default format for data representation needed for data mining tasks on the Weka environment. The default file type is called the attribute relation file format (.arff). the arff file type stores three category of data: the first defining the title of the relation, the second defining the relation's attributes alongside their respective labels and the third defining the relations data followed for the values of each attributes for each record. Also, data can be read from comma separated values (.csv) format and from databases using Object-Database Connectivity (ODBC).

4. Results and Discussion

In this section of the study, the results of the methodological approach described earlier are discussed. 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 this study were also discretized into nominal values so as to reduce the computational complexity associated with numeric variable. Following this, the results of the model formulation and simulation process for the development of the predictive model for the survival of diabetes mellitus was presented.

4.1. Result of Data Identification and Collection

The analysis of the data containing information about the attributes for the 29 patients are shown in Tables 2 and 3. 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

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
BMI Class	Missing	3	10.4
	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 metres)	1.3	1.8	1.64	0.097
BMI	17.75	40.44	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

(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 showed that information regarding the variables used to monitor the survival of diabetes mellitus patients showed 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 results of the treatment 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 of the study further showed 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 the study 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. The results further showed that the minimum and maximum weight were 30 and 93 kg while the minimum and maximum heights were 1.3 and 1.8 metres respectively. The results further showed 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 further showed 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 3 shows a diagram of the arff file for the new training data stored in the file *diabetes Training Data .arff.*

4.2. Results of Model Formulation and Simulation

Support vector machines algorithm was used to formulate the predictive model for the survival of diabetes mellitus. SVM was 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 support vector machine algorithm was implemented using the *SMOreg* algorithm which was 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.

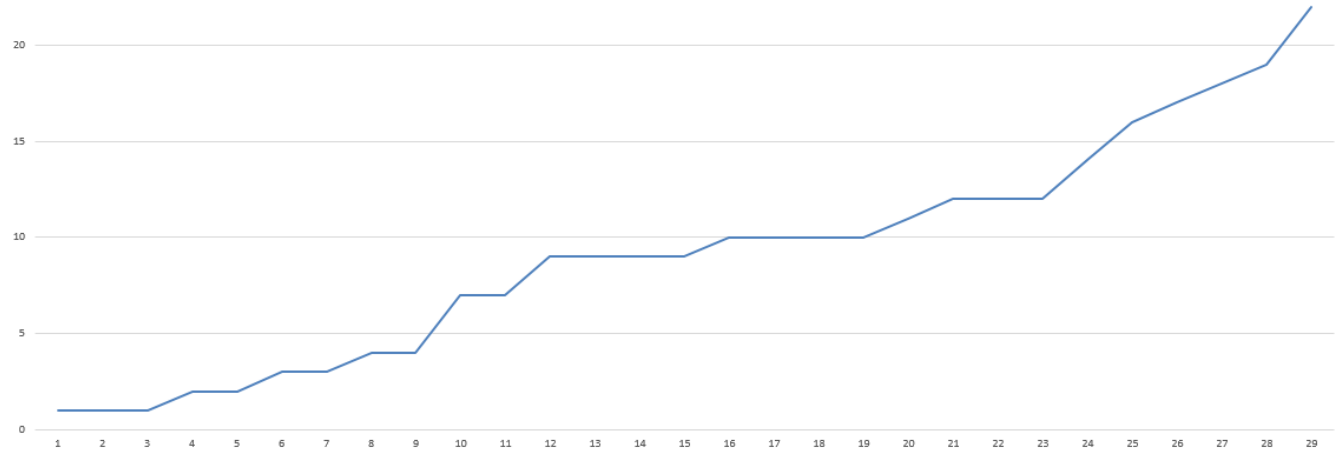

Figure 2. *Graphical plot of survival time (in years) for all patients.*

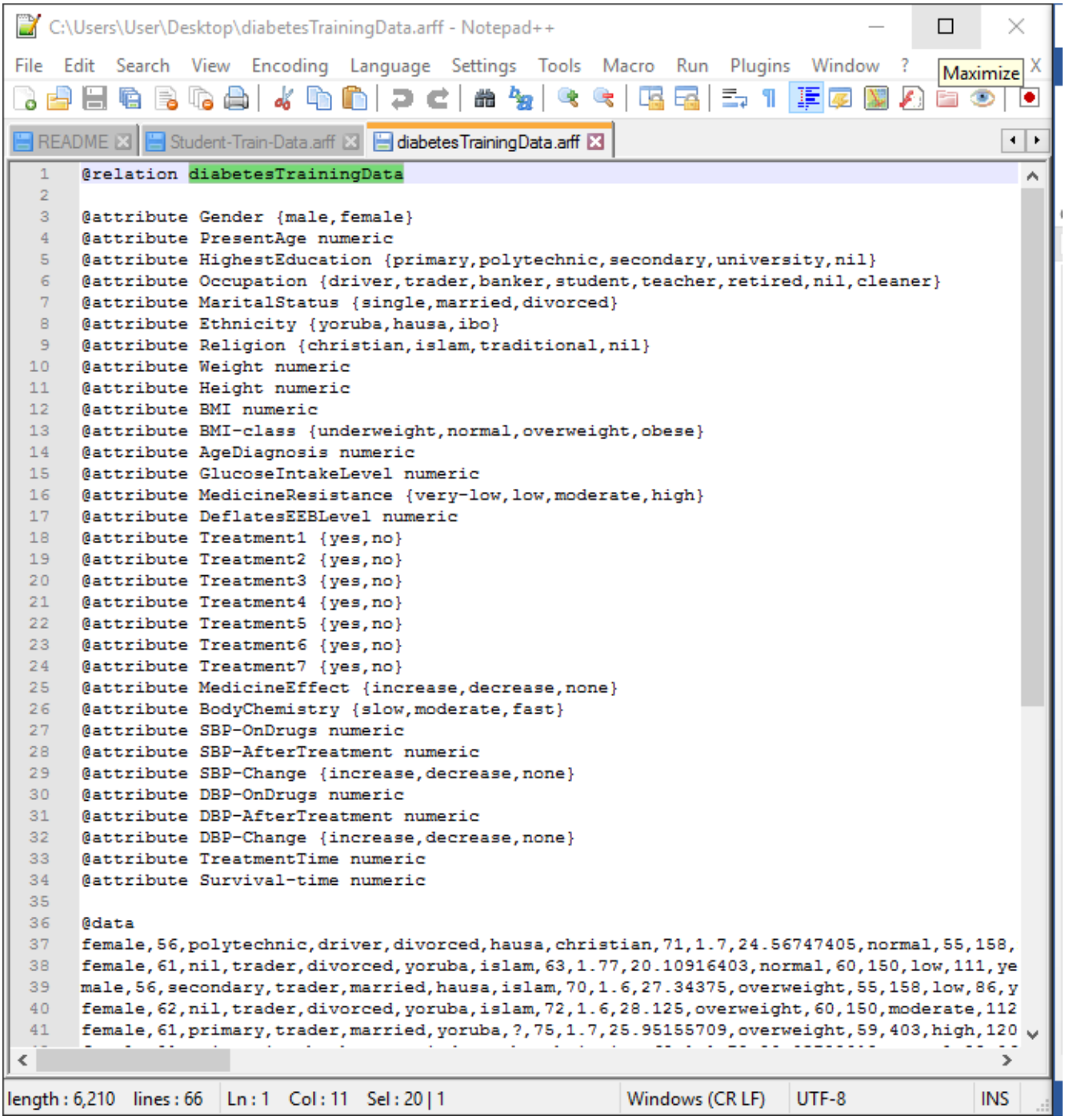

Figure 3. *Arff file containing identified attributes after data pre-processing.*

4.3. Discussion

Following the simulation of the predictive model for the survival 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 4 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 this study. The figures shows the correct and incorrect classifications made by the algorithm. Table 4 shows the distribution of the results of the actual and predicted values alongside the error of the support vector machine in determine the survival of the diabetes mellitus patients.

Figure 5 shows the graphical plot of the actual and predicted values of the support vector machines while figure 6 shows a graphical plot of the error values of each prediction made by the support vector machine algorithm. The results of the study further showed 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.

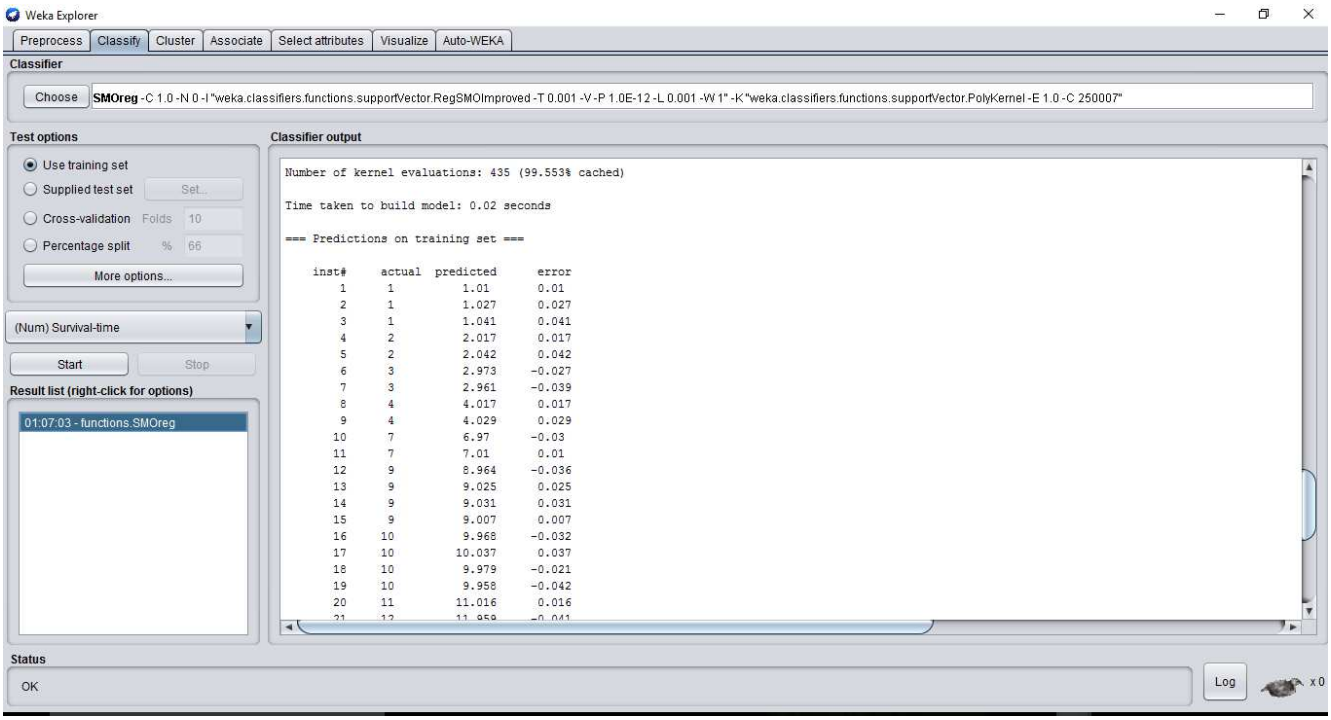

Figure 4. *Screenshot of support vector machines results on dataset.*

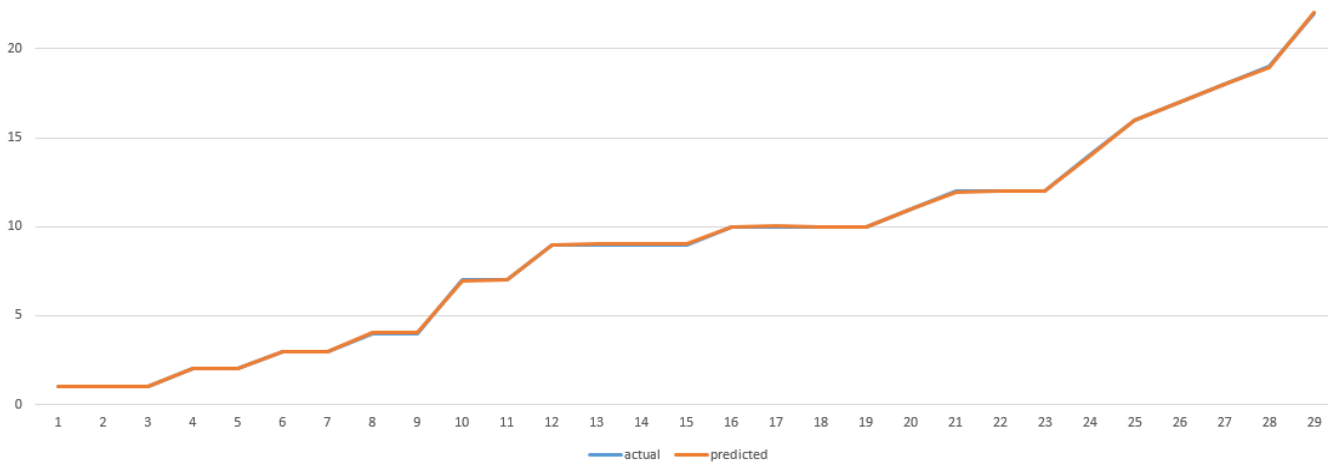

Figure 5. *Graphical plot of the actual and predicted values of the SVM model.*

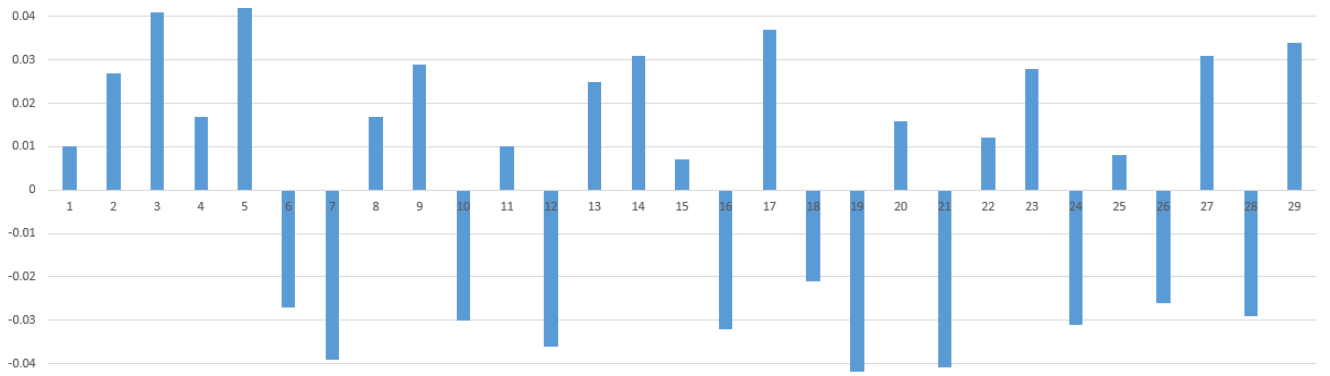

Figure 6. *Graphical plot of the error values of the SVM model.*

Table 4. *Results of the Prediction made by the SVM model.*

Instance Number	actual	Predicted	error
1	1	1.01	0.01
2	1	1.027	0.027
3	1	1.041	0.041
4	2	2.017	0.017
5	2	2.042	0.042
6	3	2.973	-0.027
7	3	2.961	-0.039
8	4	4.017	0.017
9	4	4.029	0.029
10	7	6.97	-0.03
11	7	7.01	0.01
12	9	8.964	-0.036
13	9	9.025	0.025
14	9	9.031	0.031
15	9	9.007	0.007
16	10	9.968	-0.032
17	10	10.037	0.037
18	10	9.979	-0.021
19	10	9.958	-0.042
20	11	11.016	0.016
21	12	11.959	-0.041
22	12	12.012	0.012
23	12	12.028	0.028
24	14	13.969	-0.031
25	16	16.008	0.008
26	17	16.974	-0.026
27	18	18.031	0.031
28	19	18.971	-0.029
29	22	22.034	0.034

The result of the performance evaluation of the machine learning algorithms which presents the values of the performance evaluation metrics used to evaluate the performance of the supervised machine learning algorithms selected for this study. 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.

5. Conclusion

In this paper, the development of a predictive model for predicting the survival of diabetes mellitus given the values of variables was developed using dataset collected from patients in a hospital in the south-western part of Nigeria. 32 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 following the identification of the required variables.

After the process of data collection and pre-processing, two supervised machine learning algorithms were used to develop the predictive model for the survival of diabetes mellitus using the historical dataset from which the training and testing dataset was collected. The 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

References

[1] 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.

[2] 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.

[3] Chijioke, A., Adamu, A. N. and Makusidi, A. M. (2010). Mortality pattern among type 2 diabetes patients in Ilorin, Nigeria. *JEMDSA 15* (2): 1-4.

[4] Cox, David R (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society*, Series B. 34 (2): 187–220.

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

[6] 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.

[7] 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.

[8] International Diabetes Federation Editorial Team. Mortality (2003). In: Guariguata, L., Nolan, T., Beagley, J., Linnenkamp, U. and Jacqmain, O. (Eds.). *Diabetes Atlas*, 6th edition. Brussels: International Diabetes Federation (IDF): 49.

[9] 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.

[10] Li, J., Serpen, G., Selman, S., Franchetti, M., Riesen, M. and Schneider, C. (2010). Bayes Net Classifiers for Prediction of Renal Graft Status and Survival Period. *World Academy of Science, Engineering and Technology 4* (3): 128-133.

[11] Onen, C. (1998). Diabetes morbidity and mortality in Botswana: a retrospective analysis of hospital based data on diabetic patients, 1980–1994. *International Diabetes Digest 13*: 96–9.

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

[13] 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.

[14] Waijee, A. K., Joyce, J. C. and Wang, S. J. (2010). Algorithms outperform metabolite tests in predicting response of patients with inflammatory bone disease to thiopurines. *Clin Gastroenterol Hepatol +8*: 143-150.

[15] WHO (2020). Diabetes: Key Facts. Available from https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed may 29, 2020].