Early Diabetic Risk Prediction using Machine Learning Classification Techniques

Adetunji Olusogo Julius¹, Ayeni Olusola Ayokunle², Fasanya Olawale Ibrahim³ ^{1,3} Department of Computer Engineering, Ajayi Polytechnic Ikere- Ekiti, Ekiti State, Nigeria. ² Department of Computer Science, Federal Polytechnic Ado Ekiti, Ekiti State, Nigeria.

Abstract:- Diabetes is a metabolic disorder that results from deficiency of the insulin secretion to control high sugar contents in the body system. At early stage, diabetes can be managed and controlled. Prolong diabetes leads to complication disorders such as diabetes retinopathy, angina, heart attack, stroke, atherosclerosis and even death. Therefore, assessment of diabetic risk prediction is necessary at early stage by using machine learning classification techniques based on observed sample features. The dataset used for this paper was obtained from Irvine (UCI) repository of machine learning databases and was analyzed on WEKA application platform. The dataset contains 520 samples with 17 distinct attributes. Machine learning algorithms used as classifier are K-Nearest Neighbors algorithm (KNN), Support Vector Machine (SVM), Functional Tree (FT). The evaluated results were based on parameters such as accuracy, specificity and precision. KNN has the highest accuracy, specificity and precision of 98.08%, 99% and 99.36% respectively.

Keywords: - Machine Learning Algorithm, WEKA, Diabetes, KNN, SVM, FT.

I. INTRODUCTION

Diabetes poses threat to human lives and attributed by hyperglycemia (Zou et al., 2008). Diabetes is one of the chronic disorders resulted from deficiency of insulin secretion to control high sugar contents in the body system and it affects both sexes worldwide (Mahboob et al., 2018). Based on World Health Organization (WHO) 2021 report, four hundred and twenty two million (422 000 000) people are diabetic patients worldwide with one million and six hundred thousand deaths each year, most of which are from underdeveloped or developing countries. Diabetes is not only affected by the major factor of "high sugar concentration" in the body system but with some other factors such as obesity, partial paresis, age, delayed healing, muscle stiffness, hereditary factors and deficiency in insulin secretion (Deepti and Dilip, 2018). However, early risk prediction of diabetes is a remedy from its complications (Vijayan and Anjavili, 2015). Early risk prediction of diabetes would safe medical practitioners from wasting their time and energy from clinical diagnosis of diabetic patients (Muhammad et al., 2018).

The dataset is obtained from the research paper (Faniqul Islam *et al.*, 2020). The dataset contains 520 samples with 17 distinct attributes.

II. RELATED WORK

Sakshi Gujral *et al.* (2017) adopted the application of classification techniques namely Support Vector Machine and Decision Trees for the prediction of diabetes mellitus. The dataset employed for this paper was obtained from PIMA Indian Diabetes Data-set. PIMA India is concerned with women's health. Support vector machine has the higher accuracy of 82%.

Deepti Sisodia *et al.*(2018) employed support vector machine, decision tree and naïve bayes algorithms. The research was performed on the Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The results obtained showed that naïve bayes outperforms other algorithms with the accuracy of 76.30% comparatively other algorithm.

Sneha and Gangil (2019) applied decision tree, Naïve Bayesian and random forest algorithms for the early prediction of diabetes mellitus using optimal features selection. The proposed method focused on selecting the attributes for the early detection of diabetes using predictive analysis. Decision tree and random forest algorithms have the highest specificity of 98.20% and 98.00%, respectively while Naïve Bayesian has the best accuracy of 82.30%.

Hassan, Malaserene and Leema (2020) conducted prediction of diabetes mellitus using classification techniques like Decision Tree, KNearest Neighbors, and Support Vector Machines. It was observed that support vector machine (SVM) outperforms decision tree and KNN with highest accuracy of 90.23%.

III. METHODOLOGY

The proposed system includes Data collection, Data Pre-processing, System Architecture and System Evaluation.

3.1 Data Collection

The data set used in this paper was obtained from Irvine (UCI) repository of machine learning databases. The dataset contains **520** samples with **17** distinct attributes.

3.2Data Pre – Processing

The dataset is in csv format and is being converted to arff format which is form suitable for WEKA application.

3.3System Architecture

The figure 3. 1 depicts the architecture of the proposed system in which the early stage diabetic risk prediction dataset is classified with K-Nearest Neighbors algorithm (KNN), Random Forest (RF), Support Vector Machine (SVM) and Functional Tree (FT) classifiers.



Figure 3.1 System Architecture

3.4System Pseudo code

The pseudo code of the system is enumerated in the arrangement below:

Step 1: load Dataset of diabetes into weka application

Step 2: Pre-processing of the dataset

Step 3: Classification using support vector machine (SVM) and record the result

Step 4: Classification using K-Nearest Neighbors algorithm (KNN) and record the result

Step 5: Classification using Functional Tree (FT) algorithm and record the result.

Step 6: Evaluation of the results.

Step 1 and step 2 involved loading of the dataset into the weka application and conversion of the dataset from csv format to arff format respectively. Steps 3, 4 and 5 are the classifications using support vector machine (SVM), K-Nearest Neighbors and Functional Tree (FT) algorithms respectively. **Step 6** is the evaluation of the results based on metrics namely accuracy, specificity and precision.

3.5Metrics used for classification

i. Accuracy: The percentage of the correctly classified instances i.e. accuracy, is obtained by subtracting the percentage of incorrectly classified instances from 100. Accuracy is obtained by

Accuracy =
$$\frac{Tp + TN}{TP + TN + FN + FP} \times \frac{100}{1}\%$$
.

ii. **Specificity:** Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).

Specificity =
$$\frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

iii. Precision: Precision is calculated as the number of true positives divided by the total number of true positives and false positives.

 $Precision = \frac{True Positive}{True Positive + False Positive}$

3.6 Experimental Set Up

The experimental set up involved.

i. WEKA Software: Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms for tasks in mining of data. It implements algorithms for data pre-processing, classification, regression, clustering and association rules, visualization tools are also included.



Figure 3.2: WEKA tool Applications interface (GUI Chooser)

- **ii. Dataset:** The dataset employed in this paper was obtained from Irvine (UCI) repository of machine learning databases and the paper that generated the dataset is (Faniqul Islam *et al.*, 2020).
- **iii. K-Fold Cross Validation:** The experimental set up using weka application in this paper is made up of 10-fold cross validation. The training dataset is randomly partitioned into 10 groups, the first 9 groups are used for training the classifier and the other group was used as the dataset to test on.

IV. RESULTS

4.1 Experimental Results

Figure 4.1 demonstrates how to load the dataset into WEKA application.

Preprocess Classify Cluster Associate Select attributes Visualize							
Open file Open URL Open DB	Gen	erate		Undo	Edit	Save.	
Filter							
Choose None						Appl	y Stop
Current relation		Selected at	tribute				
Relation: diabetes_data_upload Instances: 520	Attributes: 17 Sum of weights: 520	Name: Missing:	class 0 (0%)	Distinct: 2	Type: Nominal Unique: 0 (0%)		
Attributes		No.	Label	Count	Weight		
All None Invert	Pattern	1	Positive Negative	320 200	320.0 200.0		
No. Name 2 Gender 3 Polyuria 4 Polydipsia 5 sudden weight loss 6 weakness							
7 Polyphagia 8 Genital thrush 9 visual blurring 10 Itching 11 Irritability 12 delayed healing 13 partial paresis 14 muscle stiffness 15 Alopecia 16 Obesity 17 class	v	Class: class	s (Nom)		200		Visualize All
Remove							

Figure 4.1: Loading dataset into WEKA Application

Classifier											
Choose FT -I 15 -F 0 -M 15 -W 0.0											
Test options	Classifier output										
Use training set Use training set Supplied test set Set Cross-validation Folds 10 Percentage split % 66 More options (Nom) class Start Stop	Time taken to bu === Stratified o === Summary === Correctly Classi Incorrectly Classi Kappa statistic Mean absolute en Root mean square Relative absolut Root relative so	iild model cross-vali ified Inst ssified In tror ted error te error quared err	: 0.53 se dation == ances stances or	-conds = 487 33 0.86 0.08 0.22 17.75 48.96	773 441 82 68 %	93.6538 6.3462					
Result list (right-click for options)	=== Detailed Acc	curacy By	Class ===								
11:09:45 - trees.FT	Weighted Avg. === Confusion Ma a b < c 298 22 a = 11 189 b =	TP Rate 0.931 0.945 0.937 atrix === classified = Positive = Negative	FP Rate 0.055 0.069 0.060	Precision 0.964 0.896 0.938	Recall 0.931 0.945 0.937	F-Measure 0.948 0.920 0.937	MCC 0.868 0.868 0.868	ROC Area 0.968 0.968 0.968	PRC Area 0.984 0.911 0.956	Class Positive Negative	

Figure 4.2: Application of Functional Tree (FT) Algorithm

23:10:01 - functions.LibSVM

ISSN No:-2456-2165

Choose LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\Program Files\W/eka-3-8-5" -seed 1 est options **Classifier** output lime taken to build model: U.I/ seconds O Use training set O Supplied test set Set. === Stratified cross-validation === === Summary === Cross-validation Folds 10 Correctly Classified Instances 490 94.2308 \$ O Percentage split % 66 5.7692 % Incorrectly Classified Instances 30 More options... Kappa statistic 0.8793 Mean absolute error 0.0577 Root mean squared error 0.2402 Relative absolute error 12.1846 % (Nom) class 49.371 % Root relative squared error Total Number of Instances 520 Start esult list (right-click for options) === Detailed Accuracy By Class ===



Figure 4.3: Application of Support Vector Machine (SVM) Algorithm

Classifier											
Choose IBk -K 1 -W 0 -A "weka core ne	ighboursearch LinearNNS	earch -A 1"w	reka core F	uclideanDista	nce -R first-	lasti""					
	ignood could have a second			aona o an Brota							
Test options	Classifier output										
O Use training set											
O Supplied test set Set	Time taken to bu:	ild model	.: 0.01 s∈	econds							
Cross-validation Folds 10	=== Stratified c	ross-vali	dation ==	-							
O Percentage split % 66	=== Summary ===										
(Here services	Correctly Classi:	fied Inst	ances	510		98.0769	olo				
More options	Incorrectly Class	sified In	stances	10		1.9231	00				
	Kappa statistic			0.95	596						
22 - 22 - 22 - 22 - 22 - 22 - 22 - 22	Mean absolute er:	Mean absolute error		0.0207							
(Nom) class	Root mean squared	Root mean squared error		0.1388							
	Relative absolute	Relative absolute error		4.3741 %							
Start Stop	Root relative squ	uared err	or	28.52	299 %						
Result list (right-click for options)	Total Number of 3	Instances		520							
23:31:06 - lazy.IBk	=== Detailed Acc	uracy By	Class ===								
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.975	0.010	0.994	0.975	0.984	0.960	0.984	0.986	Positive	
		0.990	0.025	0.961	0.990	0.975	0.960	0.984	0.965	Negative	
	Weighted Avg.	0.981	0.016	0.981	0.981	0.981	0.960	0.984	0.978		
	=== Confusion Mat	trix ===									
	a b < c	lassified	as								
	312 8 8 =	Positive									
	2 198 b =	Negative									
	CONTRACT NO. CO. NO.	1.11.1 P. 1.1.1 P. 1.1									

Figure 4.4: Application of K-Nearest Neighbors Algorithm

Test options	Classifier output	Classifier output									
O Use training set			_		_		_				
O Supplied test set Set	Time taken to bu	ild model	.: 0.34 <mark>s</mark> e	conds							
Cross-validation Folds 10	=== Stratified c	=== Stratified cross-validation ===									
O Percentage split % 66	=== Summary ===										
	Correctly Classi	fied Inst	ances	507		97.5	\$				
More options	Incorrectly Clas	sified In	stances	13		2.5	dia i				
	Kappa statistic			0.94	72						
an a	Mean absolute er	ror		0.0566							
Nom) class	Root mean square	d error		0.13	98						
	Relative absolut	e error		11.96	12 %						
Start Stop	Root relative sq	uared err	or	28.72	85 %						
sult list (right-click for options)	Total Number of	Instances	1	520							
03:25:28 - trees.RandomForest	=== Detailed Acc	=== Detailed Accuracy By Class ===									
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
		0.978	0.030	0.981	0.978	0.980	0.947	0.998	0.998	Positive	
		0.970	0.022	0.965	0.970	0.968	0.947	0.998	0.996	Negative	
	Weighted Avg.	0.975	0.027	0.975	0.975	0.975	0.947	0.998	0.998		
	=== Confusion Ma	trix ===									
	a b < c	lassified	as								
	313 7 a =	Positive									
	6 194 b =	Negative									

Figure 4.5: Application of Random Forest Algorithm

 Table 4.1: Summary of confusion matrix and Results of evaluation metrics for the classification using machine learning algorithms

Number	Algorithms	True Positive	False Negative	False Positive	True Negative	Accuracy (%)	Precision (%)	Specificity (%)
1	K-Nearest Neighbors	312	8	2	198	98.08	99.36	99.00
2	Support Vector Machine (SVM)	300	20	10	190	94.23	96.77	95.00
3	Functional Tree (FT)	298	22	11	189	93.65	96.44	94.50
4	Random Forest (RF)	313	7	6	194	97.50	98.12	97.00

V. CONCLUSION

The research work was conducted using support vector machine (SVM), K-Nearest Neighbors (KNN), Functional Tree (FT) and Random Forest (RF) Algorithms as classifiers in which K-Nearest Neighbors perform better in terms of accuracy, precision and specificity. The research work produced highest accuracy, specificity and precision through K-Nearest Neighbors algorithm.

FUTURE WORK

Further studies can be carried out using other classification techniques, adoption of feature extraction/ feature selection techniques or using other datasets.

REFERENCES

- [1]. Deepti, S. and Dilip S. (2018). Prediction of Diabetes using Classification Algorithms. International Conference on Computational Intelligence and Data Science.
- [2]. Faniqul Islam M.M et al. (2020) 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 113-125.
- [3]. Hassan A.S *et al.* (2020). Diabetes Mellitus Prediction using Classification Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5.
- [4]. Mahboob T. A *et al.* (2018). A model for early prediction of diabetes. Informatics in Medicine Unlocked.
- [5]. Muhammad A. *et al.*(2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. Proceedings of the 24th International Conference on Automation & Computing, Newcastle University.
- [6]. Sakshi Gujral *et al.* (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women.International Journal of Advanced Research in Computer Science Volume 8, No. 5.
- [7]. Naveen K.G *et al.* (2020). Prediction of Diabetes Using Machine Learning Classification Algorithms. International Journal of Scientific and Technology Research.
- [8]. Vijayan, V.V and Anjavili, C. (2015). Prediction and diagnosis of diabetes mellitus. A machine learning approach IEEE Recent advances in Intelligent Computational Systems.
- [9]. Zou *et al.*, (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in genetics.