Prediction of Diabetes Screening by Using Data Mining Algorithms

Aberham Tadese
Rift valley university, School of post graduate studies, Department of Computer science
P.O.Box 80734, Addis Ababa, Ethiopia
E-mail:-abrehamt373@gmail.com

Abstract

Diabetes is one of the most common non-communicable diseases in the world. Diabetes affects the ability to produce the hormone insulin. Thus, complications may occur if diabetes remains untreated and unidentified. That features a significant contribution to increased morbidity, mortality, and admission rates of patients in both developed and developing countries. When disease is not detected early, it leads to complications. Medical records of the cases were retrospective. Anthropometric and biochemical information was collected. From this data, four ML classification algorithms, including Decision Tree (J48), Naive-Bayes, PART rule induction, and JRIP, were used to prognosticate diabetes. Precision, recall, F-Measure, Receiver Operating Characteristics (ROC) scores, and the confusion matrix were calculated to determine the performance of the various algorithms. The performance was also measured by sensitivity and specificity. They have high classification accuracy and are generally comparable in predicting diabetes and free diabetes patients. Among the selected algorithms tested, the Decision Tree Classifier (J48) algorithm scored the highest accuracy and was the best predictor, with a classification accuracy of 92.74%.

Keywords: Diabetes, Data Mining, ML, J48, PART, JRIP, Naïve Bayes
DOI: 10.7176/JIEA/12-1-02
Publication date: March 31st 2022

I. INTRODUCTION

Diabetes mellitus (DM) is a syndrome characterized by chronic hyperglycaemia, due to an absolute or relative deficiency of circulating insulin [1]. There are three main types of diabetes: Type 1, Type 2 & Gestational diabetes. People with type 1 diabetes produce very little or no insulin at all and it is called insulin- dependent. Type 2 diabetes used to be called non-insulin- dependent diabetes or adult-onset diabetes, and accounts for at least 90% of all cases of diabetes. Gestational diabetes mellitus (GDM) is a type of diabetes characterized by high blood glucose levels during pregnancy [2].

As a result, the figure is expected to rise to 366 million by 2030.DM is the commonest of all metabolic diseases everywhere on the planet [3]. The burden of diabetes is increasing worldwide, including in developing countries like Ethiopia. The International Diabetes Federation Association reported Ethiopia to be ranked 3rd in Africa with 1.4 million DM and a prevalence of 3.32 by the year 2012 [4].

Diabetes affects all segments of the population, regardless of age and sex [5]. Diabetes of all kinds can cause complications that will increase the general risk of dying prematurely. Possible complications include attacks, strokes, renal failure, leg amputations, vision loss, and nerve damage. Poorly controlled diabetes in pregnancy increases the danger of fetal death and other complications [6].

Early stage diagnosis Diabetes mellitus may present with characteristic symptoms such as thirst, polyuria, blurring of vision, and weight loss in the absence of effective treatment; in the most severe forms, ketoacidosis or a non-kenotic hyperosmolar state may develop and cause stupor, coma, and, in the absence of effective treatment, death [7].

II. STATEMENT OF THE PROBLEM

Diabetes mellitus is a global public health issue that contributes significantly to heart disease, stroke, chronic kidney failure, leg amputation, foot ulcer, nerve damage, and eye damage. Ethiopia ranks first among the top four African countries, with 2.6 million people. This is due to number of factors, such as lack of awareness, limitation in screening protocols, less propaganda for intervention programs, globalization, rapid adaptation to western lifestyle, unhealthy eating habits like skipping the breakfast (or) eating junk foods because of financial hardships, and poor accessibility to health care services (scarcity of specialists, practitioners, health facilities, less expenditure) on diabetes or free diabetes in Ethiopia.

Diabetes prevalence is rising at an alarming rate. According to WHO fact shit, 77% of individuals with diabetes sleep in low- and middle-income countries, making socially disadvantaged countries the most vulnerable to diabetes disease-related complications. Developing countries, the human and financial costs of diabetes management are high and escalating from time to time.

According to the literature, detecting diabetes's 80% disease progression is based on clinical suspicion and is confirmed by performing a laboratory assessment of the patient's blood sample's oral glucose or sugar level. These methods aren't feasible for screening, because they require skilled manpower and are time consuming, making them not accessible to all segments of the population. According to CSA data from 1997, 84% of people in the country live in the country, and health institutions are heavily concentrated in the city's core. On the other hand, the current health care setup is a busy outpatient setting. The shortage of highly trained health care providers is an acute problem in Ethiopia. Today, this ultimately raises the cost of patient health care service for treating non-communicable diseases like diabetic patients and improves the standard of care [8].

III. RELATED WORK AND LITERATE REVIEW

In an investigation directed by Y, Hongmei, three strategies for data mining were the examination portrayed the advancement of a clinical choice organization to anticipate the presence of myocardial infraction during an associate of 4,770 patients giving intense agony at two college emergency clinics and four local area hospitals. The clinical choice organization had comparative affectability (88.0% versus 87.8%) yet a significantly higher specificity (74% versus 71%) in foreseeing the shortfall of myocardial infarction in contrast with physicians' choices if the patients needed to be conceded to the coronary consideration unit. In the event that the choice to concede depended entirely on the choice organization, the affirmation of patients without dead tissue to the coronary consideration unit would be diminished by 11.5% without antagonistically influencing patient results or the nature of care [9].

Beck, Huain, and Y. huajo concentrated on diabetes-related difficulty avoidance. Around 30 to 80 % of type 2 diabetic cases stay undiscovered. It is proposed that information be prepared; utilizing decision trees, type 2, with various levels of pervasiveness within the limelight. It has been perceived by an asymptomatic stage between the beginning of diabetic hyperglycemia and clinical conclusion within 4–7 years. During the information collection period from 2009 to 2011, techniques for gathering information from guests were regarded as a risk factor. The attribute selected individual and epidemiological linkage when patient status in light of the hour of visit facility. of choice quality are heftiness or overweight, history of diabetes in first-degree relationships, hypertension in pregnancy, privies history of gestational diabetes, history of early termination, stillbirth, and birth of a baby under 4 kg, and foundation of patient and epidemiological information. Features include age, sex, history of diabetes, and weight loss plan (BMI). The examiner utilized the procedure of J48 Algorithm to build up the decision tree in WEKA (3.9.5 version). The degree of model checked accuracy and precision of the model was 71.7 and 97.6 %, separately. The specialist reasoned that the created model utilizing the decision tree for the screening of T2DM does not need lab tests for analysis [10].

Razak and Bakar have led investigations that have some expertise in mining affiliation rules from asthma patients' profile datasets. The purpose of the examination is to identify ascribed factors that influence asthma patients. The asthma patient profile dataset during this investigation comprises of 16,384 records and 118 factors in several organizations. These attributions are assembled into segment attribute and asthma-related attribute. The mining strategy utilized includes an information readiness stage and an affiliation rules mining stage. Understanding the personality of the dataset, distinguishing information types and configurations, recognizing deficient information, breaking down information conveyance, and discretizing information are the many stages required to efficiently preprocess the information. Due to information preprocessing and purging, just 31 attribute are left to urge affiliation rules. The affiliation rules mining stage utilizes deduced algorithms. Deciding, preparing, and testing datasets, deciding limit esteems, mining affiliation rules, and affiliation rules examination are included during the execution of quality mining [11].

Bezahegn Zerihun [12] conducted a research study to develop a predictive model for pre-diabetes screening by using processing technology from Adare General Hospital 4529 diabetic instances with sixteen attributes at Hawassa City in Ethiopia for the diagnosis of pre diabetes yes or no. He focused on the implementation of the J48 decision tree and PART to affect the problem. The experiment results show that PART rules outperformed decision tree classifiers with 96.9% accuracy.

Selam, A led a task force on the measles outbreak in Ethiopia's various districts. The philosophy for building a prescient model utilizing information handling methods for this exploration was a cross-bred six-venture Cios KDP. It had six fundamental advances. Model form by 13 selected attribute for creating a foresight model. Examiner tests are directed by two arrangement algorithms, the decision tree and the naive Bayes Models, which differ in the order of a few flare-ups. The classifier has an affectability of 86.8%, indicating that the model is capable of perceiving truth, and an explicitness of 99.7%.The next analysis utilized 9 attribute and scored the simplest exactness of 93.31% with a 70% split test alternative from the contrary trials. Examination number three scored the principal precision with both test alternatives. The Chosen algorithm suggests a district-based measles episode forecast [13].

The Shegaw-led study [14] has some expertise within the research on the expected appropriateness of data mining innovation to predict child mortality based on side-by side comparison of local area-based epidemiological datasets. The analyst utilized neural organization and selection tree strategies. Assembling and testing the models utilizing the neural organization approach, the least difficult model was distinguished for the preparation it made by utilizing the default boundaries of 9 attributes. This model had a precision pace of 93%. This classifier happened with an exactness of 95% in preparing cases, and it accomplished 95% exactness in experiments.

Aiswarya Iyer et al [15] have utilized two methods, to be specific, the Decision Tree and Naive Bayes algorithms for the conclusion of diabetes utilizing arrangement mining strategies from the University of California, Irvine (UCI) Pima Indians diabetes data set of public establishment of diabetes and stomach-related and kidney sicknesses, with 768 examples and eight traits with class mark tried positive and tried negative. The trial results show that the Naive Bayes calculation with 79.5652% precision outperformed the Decision Tree (J48) exactness of 76.9565% by a rate split of 70:30.

IV.METHODOLOGY

A. Research Design

This study follows an experimental research approach. This is because experiments that will occur to extract results from real-world implementations will be and it is important to reiterate that all the experiments and results should be reproducible. The CRISP-DM technique is followed to explore the utility of knowledge mining in diabetes screening across all eligible groups. This model was chosen since it exhibits all the benefits of the well-known and widely used methodology called CRISP-DM and provides a more general, research-oriented description. Data processing technology provides a user-oriented approach to novel and hidden patterns within the data.

B. Source of population

All DM-diagnosed diabetes medical records in Zewditu Memorial Hospital from January 2021 to March 2021

C. Data Understanding

The primary source of knowledge for this research is diabetic and free diabetes patient data from patient folders. This method does not go far enough to record missing values and fill them with inconsistencies. The researcher changed the way data collection is done. Data was collected from diabetic clinics and other general checkup and chronic follow-up units from January 2021 to March 2021 and used for building the model. The collected data was in a patient folder. It contains a complete set of 731 records about patients from Zeweditu Memorial Hospital. The dataset contains both numeric and nominal values.

D. Data Collection Methods

For studies, primary and secondary data collection methods are used as sources of data. The first data was gathered by using interviews with domain experts, and therefore the second data was gathered from different written documents, conference articles, and journal publications. A dataset was collected from baseline diabetic patients' medical history using a secondary data collection method, also referred to as the retrospective method.

E. Evaluation

Evaluate the performance and accuracy of the model created by the J48 decision tree, Naive Bayes, JRIP, and PART rule induction. The methods' relevance was checked using a confusion matrix, ROC curve, 10 folds cross validation, and a ready dataset spited with 70% split for training and 30% for testing.

V. DATA UNDERSTANDING AND PREPARATION

1. Handling Missing Value

There were some missing values in the data collected for this research project, such as the type of food typically consumed and the age of the patient. This is often corrected by the time of the next visit, and a few of them, with the assistance of the domain expert's special sorts of diabetes support characterization risk factor of the patient, in order that all the missing values are crammed with the acceptable value.

No	Attribute	Missed Values
1	Pregnancies	108
2	Glucose	5
3	Blood Pressure	35
4	Skin Thickness	216
5	Insulin	354
6	BMI	11
7	Cholesterol	1
8	DBP	18

Table 5. 1 Attributes with missing values

2. Data Discretization

Interval labels are often used to replace actual data values. For instance, smoothing techniques, including binning and dividing value by hierarchal derived new attribute construction, are the most commonly used ones. From the dataset, the "AGE" and "BMI" attributes are continuous value changes to discrete value thoughts in a discretized (binned) process. After completing the discretization process, the distinct values of the age attribute were reduced to 6 from 46 distinct values.

No	Original Attributes	New value
1	Age of participant	0-34,35-43, 44-52,53-61,62-70, >71
2	Body mass index(BMI)	BMI <=11.8 under weight, BMI =11.8-22.36kg/m2 = Normal, BMI =23-33.6 g/m2= overweight, BMI = 34-44.7 kg/m2 = obese Class1, BMI = 45-55.9 kg/m2= obese Class2, BMI >= 56 kg/m2 = very obese.

Table 5. 2 Summary of Derived Attributed with Their Values

VI. EXPERIMENTATION AND RESULTS ANALYSIS

1. J48 Algorithm

Experiment I:

This experiment was conducted under the 10-fold cross-validation test option with default parameters of Weka and the algorithm generates a model as a decision tree with 91 leaves and a size of 176. The correctly classified instances were 467, which means 63.88%, and the incorrectly classified instances were 264, which means 36.11% of the total number of instances of 731, taking 0.01 seconds to build the model.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
J48	10-fold	52.2%	51.4%	59.4%	Diabetes
		70.08%	71.4%	59.4%	Free Diabetes

Table 6. 1 10-fold test for J48 algorithm

Experiment II:

This experiment was conducted using the percentage split test option to train and test the classification model. Out of the 731 total records, 219 (70%) of the instances were used as a training dataset and the remaining 512 (30%) of the instances were used as a testing dataset. The J48 learning algorithm scored an accuracy of 138 out of 219 total testing instances. 138 (63.01%) of them were classified correctly, and the remaining 81 (36.98%) testing instances were incorrectly classified. The algorithm generates a model as a discussion tree with 91 types of leaves and 176 sizes of the tree and takes 0.06 seconds to build the model.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
Naïve Bayes	70 % split	51.2%	50.6%	59%	Diabetes
		70.1%	70.6%	59%	Free Diabetes

Table 6. 2 70% split test for J48 Classification algorithm

To conclude, the above two experiments, namely experiments I and II, were performed in order to build the classification model using the J48 classification algorithm by applying k-fold cross validation and percentage split methods, respectively, to the experiments.

Detailed Accuracy by Class				
J48	Precision	Recall	ROC Area	Class
	92.4%	88%	96.8%	Diabetes
	92.9%	95.6%	96.8%	Free diabetes

Table 5. 3 Detailed Accuracy by Class for J48 classification algorithm

Confusion matrix for J48 Algorithm

The confusion matrix may be useful for analyzing how well the classifier can recognize tuples of various classes. The two-way table's sensitivity (true positive rate) is (243/(243+33))*100 = 88.04%, and the specificity (true negative rate) of support vector machine experiments is (435/(435+20))*100 = 95.60%.The overall accuracy of this training algorithm was 91.82%, which is significantly lower than the other two algorithms used in this study.

Confusion Matrix		
Diabetes	Free diabetes	Class
243	33	Diabetes
20	435	Free diabetes

Table 6.4 Confusion Matrix for J48 Decision Tree algorithm

ROC Analysis for J48 Algorithm

ROC analysis provides tools to pick the simplest models and discard suboptimal ones. Because of the cost-benefit analysis of diagnostic decision, ROC analysis is said during a street. Figure 6.1 depicts the world under ROC for diabetes screening cases. Out came that yes, gives the ROC accuracy of 98.45% of algorithms selected from all 18 attributes.

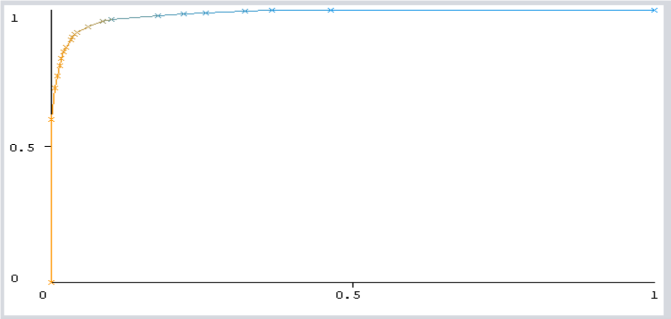

Figure 6. 1 ROC curve of the J48 classification algorithm

2. PART Algorithm

Experiment I:

This experiment was conducted under the 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as PART and correctly classified instances are 458, which means 62.65 % and incorrectly classified instances are 273, which means 37.34% of the total number of 731 instances and it takes 0.03 seconds to build the model.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
PART	10-fold	50.6%	49.3%	63.3%	Diabetes
		69.7%	70.8%	63.3%	Free Diabetes

Table 6. 5 10 fold test for PART Classification algorithm

Experiment II:

To train and test the classification model, use the percentage split test option. Out of the 731 total records, 219 (70%) of the instances were used as a training dataset and the remaining 512 instances (30%) were used as a testing dataset. The PART algorithm scored an accuracy of 133 out of a total of 219 testing instances. 133 (60.73%) of them were classified correctly, and the remaining 86 (39.26%) testing instances were misclassified or incorrectly classified.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
PART	70% split	48.1%	47.0%	55.9%	Diabetes
		70.04%	69.1%	55.9%	Free Diabetes

Table 6. 6 70% split test for PART Classification algorithm

Experiment I and Experiment II show the classification accuracy of the models based on the above two methods, respectively. The first experiment was performed based on the 10-fold cross validation method and classified with a 62.65% accuracy rate, and the second experiment, performed based on a 70%:30% percentage split, classified with a 60.83% accuracy rate.

Detailed Accuracy by Class				
PART	Precision	Recall	ROC Area	Class
	97.0%	81.5%	97.2%	Diabetes
	89.8%	98.5%	97.2%	Free diabetes

Table 6. 7 Detailed Accuracy by Class for PART algorithm

Confusion matrix for PART algorithm

The confusion matrix may be useful for analysing how well the classifier can recognize tuples of various classes. The two-way table's sensitivity (true positive rate) is (255/(255+51))*100 = 83.3%, and its specificity (true negative rate) is (448/(448+7))*100 = 98.46%.The overall accuracy of this training algorithm was 92.06%, which is significantly lower than the other two algorithms used in this study.

Confusion Matrix		
Diabetes	Free diabetes	Class
255	51	Diabetes
7	448	Free diabetes

Table 6. 8 Confusion Matrix for PART algorithm

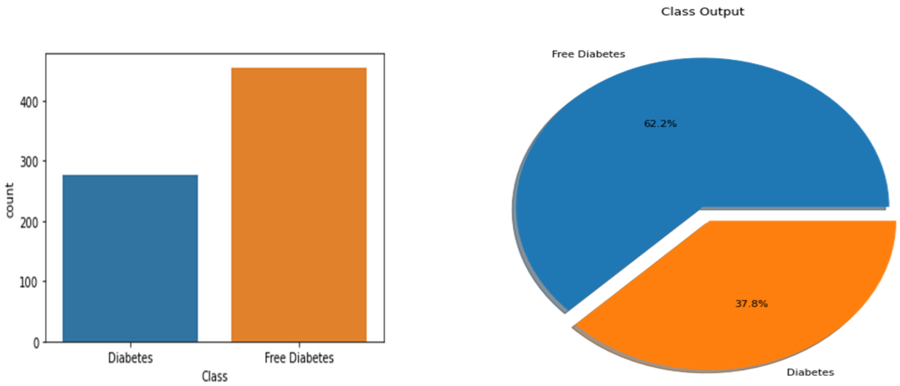

Figure 6. 2 specifying the number of people suffering by diabetes

ROC Analysis for PART Algorithm

ROC analysis is directly related to measuring the cost-benefit analysis of diagnostic PART Rule induction. Figure 6.3 shows the area under ROC for the pre diabetes screening instances. The ROC accuracy of algorithms selected from all attributes is 99.22% when class value is yes.

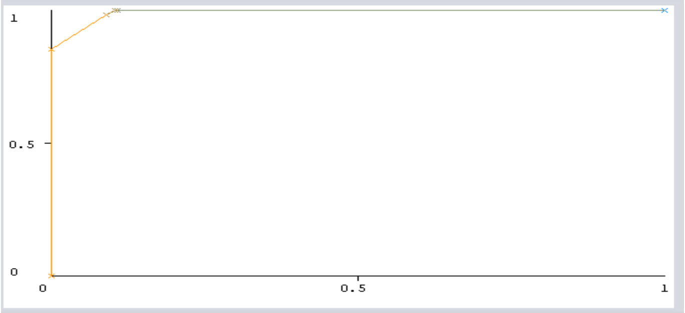

Figure 6. 3 ROC curve of the PART algorithm

3. Naive Bayes Algorithm

Experiment I:

This experiment was conducted under the 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as Naive Bayes and Correctly Classified Instances are 487, which means 66.21 % and Incorrectly Classified Instances are 247, which means 37.78% of the total number of 731 instances.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
Naive Bayes	10-fold	56.6%	45.3%	66.2%	Diabetes
		70.4%	78.9%	66.2%	Free Diabetes

Table 6. 9 10 fold test for Naive Bayes classification algorithm

Experiment II:

To train and test the classification model, use the percentage split test option. Out of the 731 total records, 219 (70%) of the instances were used as a training dataset and the remaining 512 instances (30%) were used as a testing dataset. The Naive Bayes learning algorithm scored an accuracy of out of a total of 512 testing instances, 291 (56.83%) of them were classified correctly and the remaining 221 (43.16%) testing instances were misclassified or incorrectly classified.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
Naïve Bayes	70% split	49.3%	43.4%	64.8%	Diabetes
		67.8%	72.8%	64.8%	Free Diabetes

Table 6. 10 70% split for Naïve Bayes classification algorithm

Experiment I and Experiment II show the classification accuracy of the models based on the above two methods, respectively. The first experiment was performed based on the 10-fold cross validation method and classified with a 62.21% accuracy rate, and the second experiment, performed based on a 70%:30% percentage split, classified with a 61.64% accuracy rate.

Detailed Accuracy by Class				
Naïve Bayes	Precision	Recall	ROC Area	Class
	59.7%	47.8%	70.3%	Diabetes
	71.8%	80.4%	70.3%	Free diabetes

Table 6. 11 Detailed Accuracy by Class for Naïve Bayes algorithm

Confusion matrix Naive Bayes Algorithm

The two-way table's sensitivity (true positive rate) is (132/(132+144))*100 = 47.8%, and the specificity (true negative rate) of support vector machine experiments is (336/(336+89))*100 = 79.05%.The overall accuracy of this training algorithm was 68.12%.

Confusion Matrix		
diabetes	Free diabetes	Class
132	144	Diabetes
89	336	Free diabetes

Table 6. 12 Confusion Matrix for Naïve Bayes Algorithm

ROC Analysis for Navies Bayes Algorithm

ROC analysis is performed during a cost-benefit analysis of diagnostic decisions. Figure 6.4 shows the world under ROC for diabetes screening instances. Class value of yes, gives the ROC accuracy of 70.31% of algorithms selected attributes.

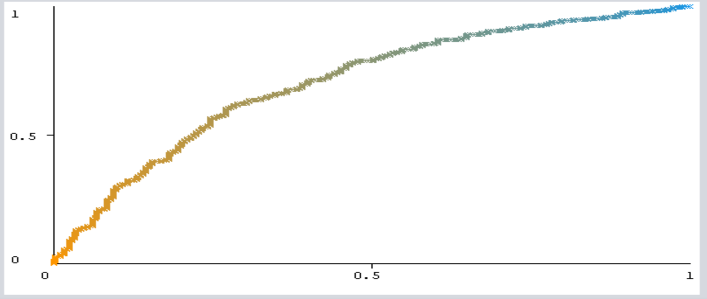

Figure 6. 4 ROC curve of the Navies Bayer's Algorithm

4. JRIP Algorithm

Experiment I:

This experiment was performed using the JRIP Rule induction algorithm with 10-fold cross validation, and the outcome of this experiment is presented in table 6.13 below.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
JRIP	10-fold	56.3%	52.2%	64.6%	Diabetes
		72.4%	75.4 %	64.6%	Free Diabetes

Table 6. 13 10-fold cross validation for JRIP algorithm

Experiment II:

The JRIP algorithm scored an accuracy of out of a total of 219 testing instances, 147 (67.12%) of them were classified correctly and the remaining 72 (32.87%) were incorrectly classified.

Algorithm	Test Option	Precision	Recall	ROC Area	Class
JRIP	70% split	56.3%	59.3%	66.9%	Diabetes
		74.2%	72.1%	66.9%	Free Diabetes

Table 6. 14 70% split for JRIP classification algorithm

To conclude, the above two experiments, namely experiments I and II, were performed so as to build the classification model using the JRIP classification algorithm by applying k-fold cross validation and percentage split methods, respectively, to the experiments.

Detailed Accuracy by Class				
JRIP	Precision	Recall	ROC Area	Class
	64.5%	49.3%	66.4%	Diabetes
	73.1%	83.5%	66.4%	Free diabetes

Table 6. 15 Detailed Accuracy by Class for JRIP algorithm

Confusion Matrix for JRIP Algorithm

The two-way table's sensitivity (true positive rate) is (136/(136+140))*100 = 49.27%, and the specificity (true negative rate) of support vector machine experiments is (380/(380+75))*100 = 83.51%.The overall accuracy of this training algorithm was 70.58%, which is significantly lower than the other two algorithms used in this study.

Confusion Matrix		
diabetes	Free diabetes	Class
136	140	Diabetes
75	380	Free diabetes

Table 6. 16 Confusion Matrix for JRIP Algorithm

ROC Analysis for JRIP Algorithm

ROC analysis is performed during a cost-benefit analysis of diagnostic decisions. Figure 6.5 depicts the world under ROC for diabetes screening cases. Class value: yes, gives the ROC accuracy of 66.87% of the selected attribute.

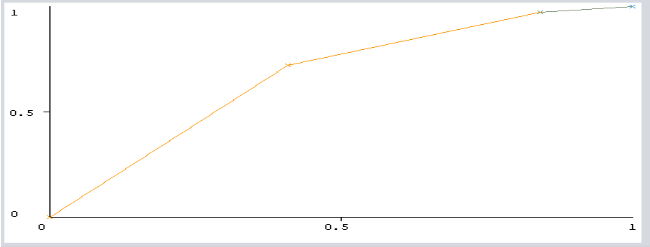

Figure 6. 5 ROC curve of the JRIP Algorithm

Comparison among Classification Algorithms

One of the aims of this research is to select a better classification Algorithm for building a model that performs best in classification. Therefore, the below table compares the output of all the four models supported by the accuracy of the model, the time it took to build the model, the sensitivity classified instances (Yes), and the insensitivity classified instances (No), supported by the 10-fold cross-validation and 70% split test option.

	10 fold test option		70% split test option	
Algorithm	Correctly classified	Incorrectly Classified	Correctly classified	Incorrectly Classified
J48	62.51%	37.48%	61.64%	38.35%
Navies Bayer’s	66.21%	33.79%	61.64%	38.35%
PART	62.38%	37.61%	59.36%	40.63%
JRIP	66.62%	33.37%	67.12%	32.87%

Table 6. 17 Comparison of 10 fold test and 70% split test option

Among the tested classification algorithms, the JRIP algorithm had the highest accuracy of 67.12%.Accordingly, this algorithm was chosen for classifications of diabetes risk.

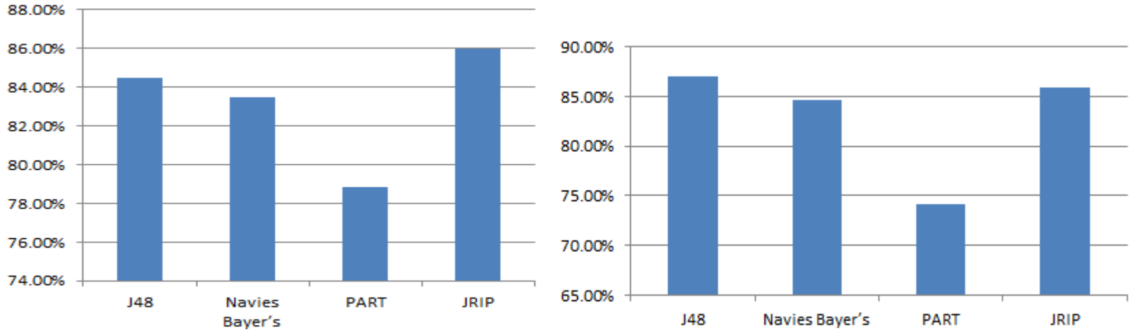

Figure 6. 6 Predicted Accuracy of each 10 fold test and 70% split Algorithm

VII. DISCUSSION RESULT ON THE MAJOR FINDINGS

For this study, the algorithms were selected to test on the diabetic datasets in order to generate rules, i.e., J48, PART, Navies Bayer’s and JRIP algorithms. Therefore, analysing one by one and seeing the result that they performed during the previous experiment has been tabularized accordingly.

The J48 algorithm is the most accurate model among the others due to the results that this algorithm demonstrated in terms of performance, time, labelling, specificity, and confusion matrix. From the previous situation, the J48 algorithm had scored a time of 0.02 seconds to classify the 678 records according to the class they belong to. Besides this, the model also showed good performance more often than others. The ROC that this model displays is almost identical to one that is 96.8 and the results of precision and recall (92.9% and 95.6%) are also pretty much the same as the left model.

The second most performing model is the PART Classier, or model which is the second one according to the above criteria for performance. This model scored the highest accuracy (92.06%) on the general data to classify the status of diabetic patient datasets. The time taken to perform the general data by this algorithm is 0 seconds, as is the time taken to classify the 673 instances of the records. The precision was 89.8%. and recall (98.5%).This result is the most promising result next to the J48 algorithm by understanding the experiment result of the model.

The third most performing model is the JRIP Algorithm model, which is the third one according to the above criteria performance, which is almost very close to the JRIP classifier. This model scored the highest accuracy (70.58%) on the general data to classify the status of diabetic patient datasets. The time taken to perform the general data by this algorithm is 0 seconds, as is the time taken to classify the 516 instances of the records. The precision was 73.1%, and the recall was 83.5%. This result is the most promising result next to the JRIP algorithm by understanding the experiment result of the model.

The fourth most performing model is the Naive Bayes Algorithm model, which is the third one according to the above criteria performance, which is almost very close to the Naive Bayes classifier. This model scored 68.12% accuracy on the general data to classify the status of diabetic patient datasets. The time taken to perform the general data by this algorithm is 0.1 seconds, as is the time taken to classify the 498 instances of the records. The precision (71.8%) and recall (80.4%).This result is the most promising result next to Naïve Bayes algorithm by understanding the experiment result of the model.

Generally, the J48 model is the most performing model with good accuracy of results. The PART rule induction is the second most performing model next to the J48 model, whereas the JRIP and Naive Bayes algorithms are the last performing classifiers. Among these algorithms, the J48 algorithm is the best performing model by classifying diabetic patient datasets and generating rules.

VIII. CONCLUSION AND RECOMMENDATIONS

Conclusion

This experimental research, which engaged a CRISP methodological approach, made use of predictive modeling techniques to address the problem. The experiment result shows the selected algorithms tested, the decision tree classifier (J48) algorithm scored the highest accuracy and best predictor with (92.74%), followed by PART (92.06%), JRIP (70.58%), and Naive Bayes algorithms (68.12%).

Recommendation and Future Work

This study showed the potential applicability of data mining algorithms to diabetes screening datasets in developing a classification algorithm model. Based on the findings of the study, we recommend the following as future research directions:

- We used the J48 decision tree, the PART, the JRIP, and the Naive Bayes classifier. Further research using ANN, KNN, SVM, and others is required to improve the performance of the predictive model.
- It is difficult to get well-organized, correct, and quality data for the mining algorithms. We suggest health centres analyse their data symmetrically for data analyses.
- More research and development efforts need to be conducted to enable and explore the variety of data mining techniques that can be applied to diabetes and free diabetic datasets.
- Integration of data mining techniques into existing systems and computerizing manual recording systems in databases is a priority issue.
- We would like to develop web-based software for performance evaluation of various classifiers where the users can just submit their data set and evaluate the results on the patient.

References

[1] A. Sinclair, P. Saeedi, A. Kaundal, S. Karuranga, B. Malanda, and R. Williams, “Diabetes and global ageing among 65–99-year-old adults: Findings from the International Diabetes Federation Diabetes Atlas, 9th edition,” *Diabetes Res. Clin. Pract.*, vol. 162, p. 108078, 2020.

[2] R. Williams *et al.*, “Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 9th edition,” *Diabetes Res. Clin. Pract.*, vol. 162, 2020.

[3] W. Gao and Q. Qiao, “Screening for type 2 diabetes,” *Epidemiol. Type 2 Diabetes*, pp. 29–38, 2012.

[4] B. S. Kumar and D. G. R., “A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis,” *Ijarcce*, vol. 5, no. 12, pp. 463–467, 2016.

[5] A. B. Adeyemo, “On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network pModels Model s,” vol. 4, no. 2, pp. 1–8, 2011.

[6] J. M. Dowling and C.-F. Yap, “Communicable Diseases in Developing Countries,” *Commun. Dis. Dev. Ctries.*, 2014.

[7] Government of India: Ministry of Health and Family Welfare, “National Guidelines for Diagnosis & Management of Viral Hepatitis,” no. December, pp. 1–88, 2018.

[8] B. Dagnew *et al.*, “Hypertriglyceridemia and Other Plasma Lipid Profile Abnormalities among People Living with Diabetes Mellitus in Ethiopia: A Systematic Review and Meta-Analysis,” *Biomed Res. Int.*, vol. 2021, 2021.

[9] P. Bernhard, K. Driessens, and P. Reutemann, “Collective and semi-supervised classification,” *Univ. Waikato, Tech. Pap.*, pp. 1–21, 2014.

[10] S. Girma, “Developing a Predictive Model to Determine Higher Education Students’ Academic Status Using Data Mining Technology.” St. Mary’s University, 2019.

[11] K. Sharma, S. Vashisht, H. Sharma, R. Dhiman, and J. K. Bains, “A Hybrid Approach Based On Association Rule Mining and Rule Induction in Data Mining.” Citeseer.

[12] B. Zerihun, “Developing a Predictive Model for Pre-Diabetes Screening by Using Data Mining Technology.”

Addis Ababa University, 2017.

[13] A. SELAM, “PREDICTING THE OCCURRENCE OF MEASLES OUTBREAK IN ETHIOPIA USING DATA MINING TECHNOLOGY.” Addis Ababa University, 2012.

[14] S. Anagaw, “Application of data mining technology to predict child mortality patterns: the case of butajira rural health project (brhp),” *Unpubl. Masters thesis Addiss Ababa Univ.*, 2002.

[15] K. P. Tripathi, “A review on knowledge-based expert system: concept and architecture,” *IJCA Spec. Issue Artif. Intell. Tech. Approaches Pract. Appl.*, vol. 4, pp. 19–23, 2011.