IoT Based Clinical Decision Support System Using Classification Technique

Submitted By

Md. Shahriar Hassan
ID: 2013-3-60-028

Atiqur Rahman
ID: 2013-3-60-012

Supervised By

Dr. Ahmed Wasif Reza
Associate Professor,
Department of CSE, EWU.

A project

Submitted in partial fulfillment of the requirements

for the degree of Bachelor of Science to

Computer Science and Engineering

Dhaka, Bangladesh

August, 2017

Declaration

This thesis has been submitted to the department of Computer Science and Engineering, East West University in the partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering by us under the supervision of Dr. Ahmed Wasif Reza, Associate Professor at Department of CSE at East West University under the course 'CSE 497'. We also declare that this thesis has not been submitted elsewhere for the requirement of any degree or any other purposes. This thesis complies with the regulations of this University and meets the accepted standards with respect to originality and quality. We hereby release this thesis to the public. We also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

_____

Md. Shahriar Hassan

ID: 2013-3-60-028

Department of Computer Science and Engineering

East West University.

_____

Atiqur Rahman

ID: 2013-3-60-012

Department of Computer Science and Engineering

East West University.

Letter of Acceptance

The thesis entitled "IoT Based Clinical Decision Support System Using Classification Technique" submitted by Md. Shahriar Hassan, ID 2013-3-60-028 & Atiqur Rahman, ID 2013-3-60-012 to the department of Computer Science & Engineering, East West University, Dhaka 1212, Bangladesh is accepted as satisfactory for partial fulfillments for the degree of Bachelor of Science in Computer Science & Engineering in August 2017.

Board of Examiners

1_____

Dr. Ahmed Wasif Reza

Associate Professor — Supervisor

Department of Computer Science and Engineering

East West University, Dhaka, Bangladesh

2_____

Dr. Md. Mozammel Huq Azad Khan

Professor and Chairperson — Chairperson

Department of Computer Science and Engineering

East West University, Dhaka, Bangladesh

Acknowledgements

First of all, we are grateful to the Almighty God for establishing me to complete this research. We wish to express our sincere thanks and gratitude to our advisor Dr. Ahmed Wasif Reza, Associate Professor at Dept. of CSE for the continuous support during our thesis study and related research, for his patience, motivation, and immense knowledge. His guidance helped us in all the time of research and writing of the thesis. We will always be grateful for having the opportunity to study under him.

We are thankful to all of our teachers, Department of CSE, East West University. We also grateful to all of our primary and secondary school teachers who are our first teachers in our life and initiator of our basic knowledge.

We would like to express our thanks to our parents and siblings for supporting us spiritually throughout writing this thesis. And we are thankful to all my friends and colleagues. And at last we again thanks to the creator Allah for everything.

Abstract

In healthcare, different data mining methods are used to mine dataset and then predict diseases using medical data with the help of many machine learning methods. Diabetic disease is spread out in the whole world comprehensively. A prosperous/advanced and skillful method is presented in this research including IoT to gain a better result from the diabetic dataset. The proposed system transmits diabetic data to the database through the cloud system using mobile or smart device or hospital management. If in the dataset has any missing value or abnormal value than the proposed intelligent system will handle it and will predict disease properly. The predicted data are also stored in the database, when users or medical management send a request by legal authentication it will give the predicted results from its system. The proposed system is evaluated using “Pima Indians Diabetes” data set. We use two words, one is RAW data set and another one is NEW data set. Raw data set refers to the “Pima Indians Diabetes Data Set” as it is and the New dataset is the manipulated dataset of the raw dataset. In the new dataset we manipulate some missing and some abnormal value using our technique. In this research, we have improved the accuracy using our technique and we have used several test environments over the research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IoT	-	Internet of Things
CDSS	-	Clinical Decision Support System
IoThNet	-	Internet of Things Health Network
KNN	-	K- Nearest Neighbor
I/O	-	Input and Output
ML	-	Machine learning
AI	-	Artificial Intelligence
DM	-	Design Methodology

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Today IoT (internet of things) is a buzz word in everywhere more precisely in the technical world. IoT mainly shows a scenario where some devices not only computer but also smart phones, television, cars or other electronic devices are connecting to a network by the internet and also have computing capability extended to objects, sensors and other items by connecting that network with minimal intervention. There are many kinds of disease like diabetes, brain cancer, brain stroke, seizures, retinopathy, Macular degeneration etc. by computational health informatics research. Using the techniques of data mining and machine learning algorithm predicts the effectiveness of medical tests, medication, and discovery of relationships among huge clinical and diagnosis data. In medical system, Doctor takes some data information by many examinations, surgical procedures, and medical test but this data are not found hidden information for an effective decision. A decision is taken on the basis of doctors' perception and his experience so that the hidden data can't find correctly whether the decision true or false. The doctor or physicians may not be capable of recognizing the disease accurately. So that we can say that medical diagnosis is a very composite process. Clinical decision support systems (CDSS) are improved health and health care by raising knowledge and person specific information at appropriate times. There are two groups of CDSS namely knowledge based CDSS and non-knowledge based CDSS. In this research, we are also going to introduce a new medical system with IoT that can collect data from a hospital or a patient and analysis the given data sets and predict the disease by existing algorithms. We have proposed a useful clinical decision support

system using machine learning algorithm named C4.5 and KNN algorithm. Data mining has used in this proposed work in discovery for predictive to make more active and accurate decision result. The target of this paper is to develop a system where any existing algorithm will give more accuracy if we use our framework.

1.2 PROBLEM STATEMENTS

We are going to propose a system where we can deal with missing and abnormal values. We also propose a generalized framework that will work for any diseases. We want our build software will make diseases prediction easier than ever.

We observe that there are several limitations like many proposed system doesn't deal with missing and abnormal data. So we found problems and challenges by reading multiple papers and by understanding their proposed model. From the problem of diseases prediction and analysis, we strongly believe that we should contribute in this area. There are millions of people over the world suffering from different diseases only for lack of early detection system. Peoples are busy with their everyday life so they don't get enough time to go for regular checkup. So, we try to introduce a smarter and easier way to consult with the doctor.

At first, in our proposed system we are going to introduce a procedure to deal with missing data. From this technique, we can find the missing values and we can replace it also. We also replace any abnormal value if it exists in the data set.

1.3 RESEARCH OBJECTIVES

I. To develop IoT based clinical decision support system.

II. To analyze disease dataset according to our technique.

III. To design a system where missing and abnormal value can be found and manipulated easily.

1.4 THESIS CONTRIBUTIONS

I. We have introduced a central system where any missing and abnormal data will be replaced by possible expected data.

II. We have shown that how our system is working with any existing algorithm.

III. We design a web and mobile based application which is working in real life.

IV. We integrate IoT with our system. This system is helping to millions of patients over the world.

1.5 THESIS ORGANIZATION

The following is an overview of the contents of the chapter that presented in this research:

Chapter 2 Chapter 2 provides an overview of the literature survey on clinical decision support system, the importance of CDSS, the introduction of IoT and IoT related previous work. It also provides the overview of the missing values and many kinds of algorithms of machine learning and data mining techniques.

Chapter 3 Chapter 3 discuss our proposed methodology, describe the whole proposed system how its work and also give the algorithm and mathematical equations related to this research.

Chapter 4 In this section, we provide the experimental result of our research using different test environments. We have shown the comparison between Raw dataset and New dataset with different test environment.

Chapter 5 Chapter 5 is the concluding chapter that describes the summary of this thesis which visualization by analysis and we also provide some recommendations for further research and future works.

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY OF EXISTING TECHNIQUE

There are many kinds of research studies have been done using artificial intelligence, machine learning, and data mining algorithm in many sectors like health care. In paper [1], the author proposed a design how to create a healthcare system in Internet of things (IoT) using many network layer system and computer software, they basically design IoT healthcare network (IoThNet) which is one of the elements of IoT healthcare. Paper [3], the author had introduced many types of wearable devices for IoT and describe their many applications in health care sectors. In reference [4], there is a wrist band (hardware device) which is connected to the internet. The advanced technology and improvement of machine technologies lead to a new era Internet of things (IoT), and IoT helps make a remote control based and smart technology in healthcare [9]. In future whole world will connect with the web or the internet, internet of things is an integrated part of next generation, it has some security issue and techniques [11] that are used to control and they are access control, hashing control, steganography, cryptography and hybrid Cryptography [17][18]. In paper [14], analyzed about IoT security and privacy feature including the requirements of security issue, treat models and attack taxonomies from the health care perspective. They also designed a wearable health care context to determine how people get facilities by economic and societies in terms of sustainability.

For handling missing values in the C4.5 algorithm in the paper [19], the scientist uses some statistical method and they are Hot Deck Imputation, Cold Deck imputation, Nearest Neighbor Imputation, Substitution and Mean Substitution. In paper [22] author work with

predictive value imputation, the distribution based imputation used by the C4.5 algorithm and reduces modeling for a classification tree. In paper [23], proposed approach only numerical values to impute the missing values that also can extend to handle categorical attributes, and finally they compare with the other factors like time, space, cost etc. In the other paper [16] they presented the comparison between K-NN, decision trees and Naïve Bayes algorithm using WEKA, Rapid miner, Tanagra, Orange and Knime tools on Indian Liver Patient Dataset. For missing data, two categories are used in the paper [15] and they introduced their proposed system how to dealing with missing values in test data sets and training data sets. In the case of predicting diabetic disease, paper [13] shows a comparison with the help of WEKA tools. An IoT based information system design proposed for indoor and outdoor use in paper [10], they had introduced a Design Methodology (DM) by which it can classify the problems and notify to the healthcare center or stuff immediately.

In paper [8], they had analyzed the diabetic dataset with support vector machine (SVM), J48 algorithms, CART, K-Nearest Neighbor for finding classifications, for K-NN theirs accuracy rate for correct classification instance was 53.39% and incorrect classification instance was 46.605%. In paper [12], authors make a diagnostic classification of diabetic nephropathy and the test dataset of type 2 diabetic patients and their result summary of serum triglyceride in Naïve Bayes algorithm was 59.57% accuracy. In paper [13], they found the solution to diagnose the disease by analyzing the pattern found in the data using classification analyzing by decision tree and Naïve Bayes algorithms. Using clinical decision support system author proposed how to predict the risk level of heart disease [20]. In the paper [21] the accuracy rate is 71.4% using the C4.5 algorithm. They also compare the performance of different types of algorithms.

2.2 SUMMARY

The part of literature review has clearly brought about the fundamental overview of the study on clinical decision support system, IoT, missing value in datasets and also the security of data and other authentications. Also, study about many kinds of algorithms from data mining and machine learning (ML). There are many challenges associated with the current trend of development procedure and can discover a new procedure for further research. By analysis the existing techniques and research, our research aims to propose an intelligent system for improving the performance of clinical decision support system using IoT by removing missing and abnormal values from datasets.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 PROPOSE SYSTEM

In this research, we use "Pima Indians Diabetes Data Set" [25] as our data set. We use our technique to manipulate this data set. We strongly believe that real life data are too much complex and there should be lots of missing value as well. So, we feel that we should design a system where we can replace the missing value as well as abnormal value. For example, zero (0) is a value but this is medically impossible for someone to have this kind of value as diastolic blood pressure. Missing value refers to an empty value, not zero. Zero may be abnormal or normal and this is depending on the situation. For example, the number of pregnancy may be zero which is normal and not missing value.

We have implemented it (our proposed system) for the diabetic disease but we are having with a generalized system. At first, we will show how this system will work for any disease then we will switch to a more precise version which is the diabetic disease.

In this proposed system, a diabetic dataset of a patient is taken as the input. This data set comes from hospital database or a smartphone user or any user connected to this system. Users can also send data via their wristband sensor or smart watch. All data will be stored in a central database. When a data comes from hospital or user it may be unstructured or raunchy and there may be some missing value also. So, we have to replace an abnormal or missing value with expected value. Our algorithm will dynamically find this kind of situations and replace an abnormal value with expected value. We also save those unstructured value in the central database

because if we have those data then we can find a pattern how the user or any other data provider provides the data. When we will get everything perfectly as we expect then we will save all the data to the central database. Finally, we have the perfect data set for classifying with any existing algorithm such as C4.5, KNN. After classifying those data we can predict that someone has diabetics or not and in general someone has any complexity or not.

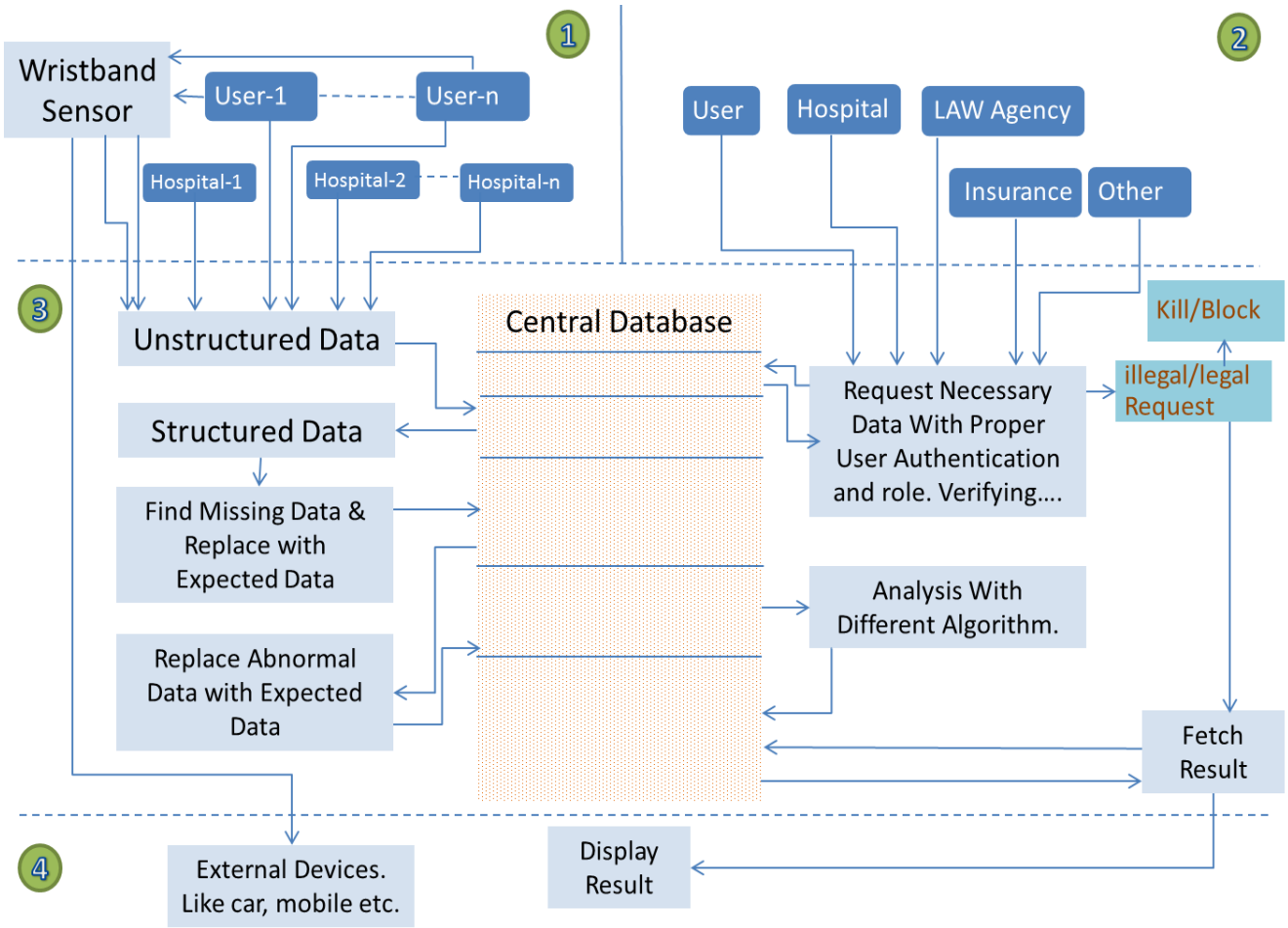


Figure 3.1: Block diagram of proposed system.

We design an expert system to predict many kinds of disease. This is our proposed system in figure 1. There are four steps in this framework. The working methodology or workflow of this system is given below.

Input Section (1): Through the wristband or smart watch or smartphone users can send attributes value of disease like plasma glucose, blood pressure etc. into a database by the internet connection (for the smartphone there is an app for use). Like the wristband hospital, Medicare also connected to the central databases and can send all the data in this central database.

Input Section (2): In this part, user/hospital/insurance etc. can see the predicted result of a patient by sending a request to the server or Management System.

Processing Section (3): This section is the important section of this proposed system. We use machine learning and data mining algorithm in this section. According to our research, we will call it "central database" where all the data which collect from the users and hospital are accumulated in this database. When users and hospitals send data, it can be unstructured data. Our system can convert those data sets into "structured data". After the arrangement, if our intelligent system found any missing or abnormal values that can be imputed by our own technique which we found with the help of statistical method. In that way, the system replaces the abnormal and missing values and stores them into the database. When there is no abnormal or missing value in the system it will start analysis all data according to C4.5, KNN or any other algorithm and predict the result for patients. The predicted result will save into the database. When the Input section (2) request into the database for showing the predicted result it will check for proper authentication and send the final result to the users or hospital.

Output Section (4): This is the output section where the result will show to user devices or any other external devices and the device will take action according to the output. For example, a user driving a car and suddenly he/she is feeling pain in his/her chest and output section will provide a signal to the car and car will stop or will do something like this and at the same time a notification will be sent to the nearest hospital and user's family or to whom the user wants.

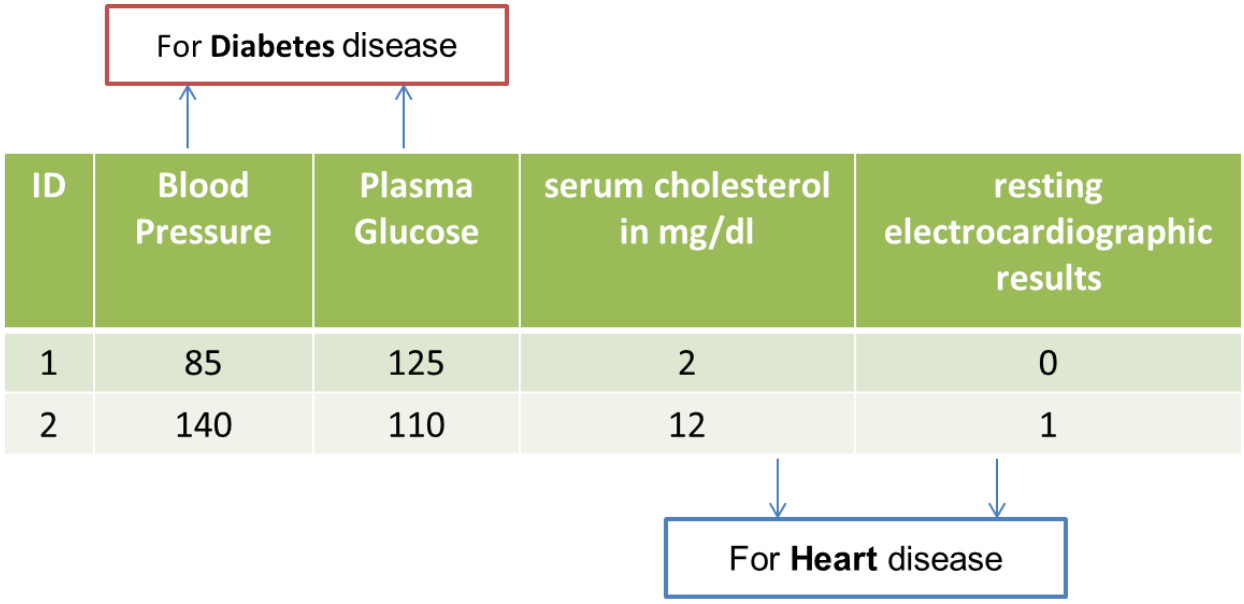


Figure – 3.2: A portion of central database.

From the figure-3.1.2 we see a portion of the central database. This is where every data will store. There will be many attributes for all the diseases and when we will work with the diabetic disease we will pick only those attribute which is related to the diabetic disease. If we work with heart disease then we will pick only heart disease related attribute. This is how we can work with any disease from this central system.

3.2 ALGORITHM AND MATHEMATICAL EQUATIONS

Step 1: data = dataset.

Step 2: pick a missing or abnormal attribute among all the attributes;

Step 3: while (missing or abnormal)

Step 3.1: if it is nominal then

Step 3.1.1: data=Different algorithm KNN (data), C4.5 (data), any other classification algorithm.

Step 3.2: else

Step 3.2.1: A=Find the mean of the specific attribute.

Step 3.2.2: B=Find the mean of those who have diabetics.

Step 3.2.3: C=Find the mean of those who have not diabetic.

Step 3.2.4: result=round ((A+B+C)/3).

Step 3.2.5: data=result.

end if

end while

Step 4: Output=Different algorithm KNN (data), C4.5 (data), any other classification algorithm

Step 5: Central database = Output

Step 6: send to user

i) For decision tree: Calculating the information gain

Gain (p) = F (Info(T) - Info(p,T)) ……..……..…….. (3.3.1)

Where, Info (T) = Entropie (p) = $-\sum_{i=1}^{n} pi \times \log(pi)$

And Info (p,T) = $\sum_{j=1}^{n} pi \times \text{Entropie}(pi)$

F= no. of unknown sample in datasets,

pi = set of probability distribution,

T= test, p = all possible values for attribute T.

Precision (confidence) $= \frac{TP}{TP+FP}$ …….…….. (3.3.2)

Recall = $\frac{TP}{TP+FN}$ …….…….. (3.3.2)

F1 measure = $2\times \frac{Precision \times Recall}{Precision + Recall}$ …….. (3.3.2)

ii) Evaluation the Performance: For overall accuracy

Accuracy = $\frac{(TP+TN)}{(TP+FP+TN+FN)} \times 100\%$ ……... (3.3.4)

Confusion matrixes mainly describe the performance of a classifier model and it contains information about actual and predicted classifications done by a classification system [24]. For evaluating the performance of such systems we have to use the data in the matrix. The following table shows the confusion matrix for a two class classifier.

- **TN** is the number of **correct** predictions that an instance is **negative**.
- **FP** is the number of **incorrect** predictions that an instance is **positive**.
- **FN** is the number of **incorrect** of predictions that an instance **negative**. and
- **TP** is the number of **correct** predictions that an instance is **positive**.

		Predicted	
		Negative	Positive
Actual	Negative	**TN**	**FP**
	Positive	**FN**	**TP**

Table 3.3: Confusion Matrix.

We need a couple of mathematical formula to calculate the accuracy, gain, precision, recall

and F1 measure. With the help of below formulas, we calculate accuracy and other staffs and those are showing in the result section.

3.3 SUMMARY

In this chapter, we discuss research methodology with a new proposed method. We have introduced our intelligent proposed system which can perform better for any machine learning algorithm. We describe how IoT is connected to this system. Also, describe the whole working flow of the proposed system in this section. For our experiment, we use the diabetic datasets, and also describe how the other disease like heart disease, cancer, jaundice etc. can be predicted by this system. All mathematical terms are briefly described in algorithms and mathematical equations section.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 EXPERIMENTAL SCENARIOS

In this research, we use two algorithms for finding accuracy and other values. At first, we use “C4.5” then “KNN” with the value of K=5. Each case we show how our algorithm gives result with any existing algorithm we use. We also use several test environments to find the accuracy in different situations. As a test environment, we use each training data as test data, once raw data as test data and new data set as training data and vice versa and cross-validation with the value of fold is 10. To figure out the accuracy we use equation 3.3.4.

Every test environment is decorated with one detailed accuracy table, one summary table, and one confusion matrix table.

I. Test Environment-1 :

- **a.** Training data: New Data Set.
- **b.** Test data: New Data Set.
- **c.** Algorithm: C4.5

Here, new data set is using for training and testing purpose. C4.5 is using as the algorithm. Table 4.1.1 is representing the detailed accuracy, table 4.1.2 is representing the summary of this test and table 4.1.3 is representing the confusing matrix by test environment-1. From the table 4.1.2 we can see that our technique successfully classify 648 instances which leads the 84.375 % accuracy.

Table - 4.1: Detailed accuracy of the new dataset for test environment-1.

			Weighted Avg.
Class	**Tested_Negative**	**Tested_Positive**	**---**
TP Rate	0.939	0.672	0.844
FP Rate	0.328	0.064	0.226
Precision	0.842	0.849	0.844
Recall	0.939	0.672	0.844
F-Measure	0.886	0.750	0.839
MCC	0.648	0.648	0.648
ROC Area	0.894	0.894	0.894
PRC Area	0.920	0.822	0.886

Table – 4.2: Summary of the new dataset for test environment-1.

	New Data
Correctly Classified Instances	648 (**84.375 %)**
Incorrectly Classified Instances	120 (15.625 %)
Kappa statistic	0.6386
Mean absolute error	0.2288
Root mean squared error	0.3382
Relative absolute error	50.3421%
Root relative squared error	70.9614%
Total Number of Instances	768

Table – 4.3: Confusion matrix of the new dataset for test environment-1.

	New Data Set	
	Predicted: No	**Predicted: Yes**
Actual: No Tested_Negative	468	32
Actual: Yes Tested_Positive	88	180

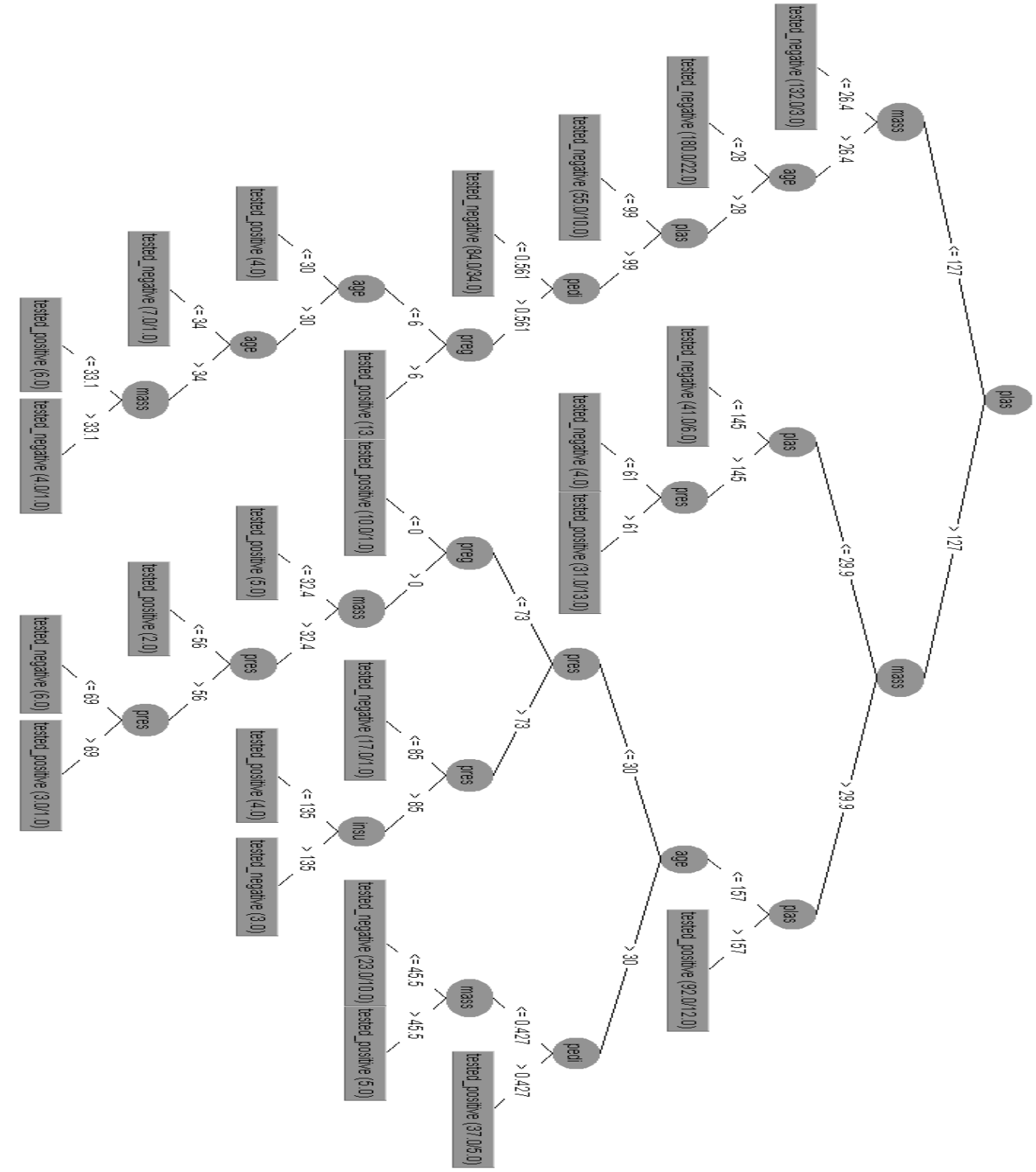


Figure - 4.1: Tree view of test environment-1.

II. Test Environment-2 :

d. Training data: New Data Set.

e. Cross-validation with 10 folds.

f. Algorithm: C4.5

Here, new data set is using for training purpose. C4.5 is using as the algorithm. Table 4.1.4 is representing the detailed accuracy, table 4.1.5 is representing the summary of this test and table 4.1.6 is representing the confusing matrix by test environment-2. From the table 4.1.5 we can see that our technique successfully classify 564 instances which leads the 73.4375 % accuracy.

Table – 4.4: Detailed accuracy of the new dataset for test environment-2.

			Weighted Avg.
Class	**Tested_Negative**	**Tested_Positive**	**---**
TP Rate	0.808	0.597	0.734
FP Rate	0.403	0.192	0.329
Precision	0.789	0.625	0.732
Recall	0.808	0.597	0.734
F-Measure	0.798	0.611	0.733
MCC	0.410	0.410	0.410
ROC Area	0.751	0.751	0.751
PRC Area	0.809	0.577	0.728

Table – 4.5: Summary of the new dataset for test environment-2.

	New Data
Correctly Classified Instances	564 (**73.4375 %)**
Incorrectly Classified Instances	204 (26.5625 %)
Kappa statistic	0.4093
Mean absolute error	0.3161
Root mean squared error	0.4477
Relative absolute error	69.541%
Root relative squared error	93.9366%
Total Number of Instances	768

Table – 4.6: Confusion matrix of the new dataset for test environment-2.

	New Data Set	
	Predicted: No	**Predicted: Yes**
Actual: No Tested_Negative	404	96
Actual: Yes Tested_Positive	108	160

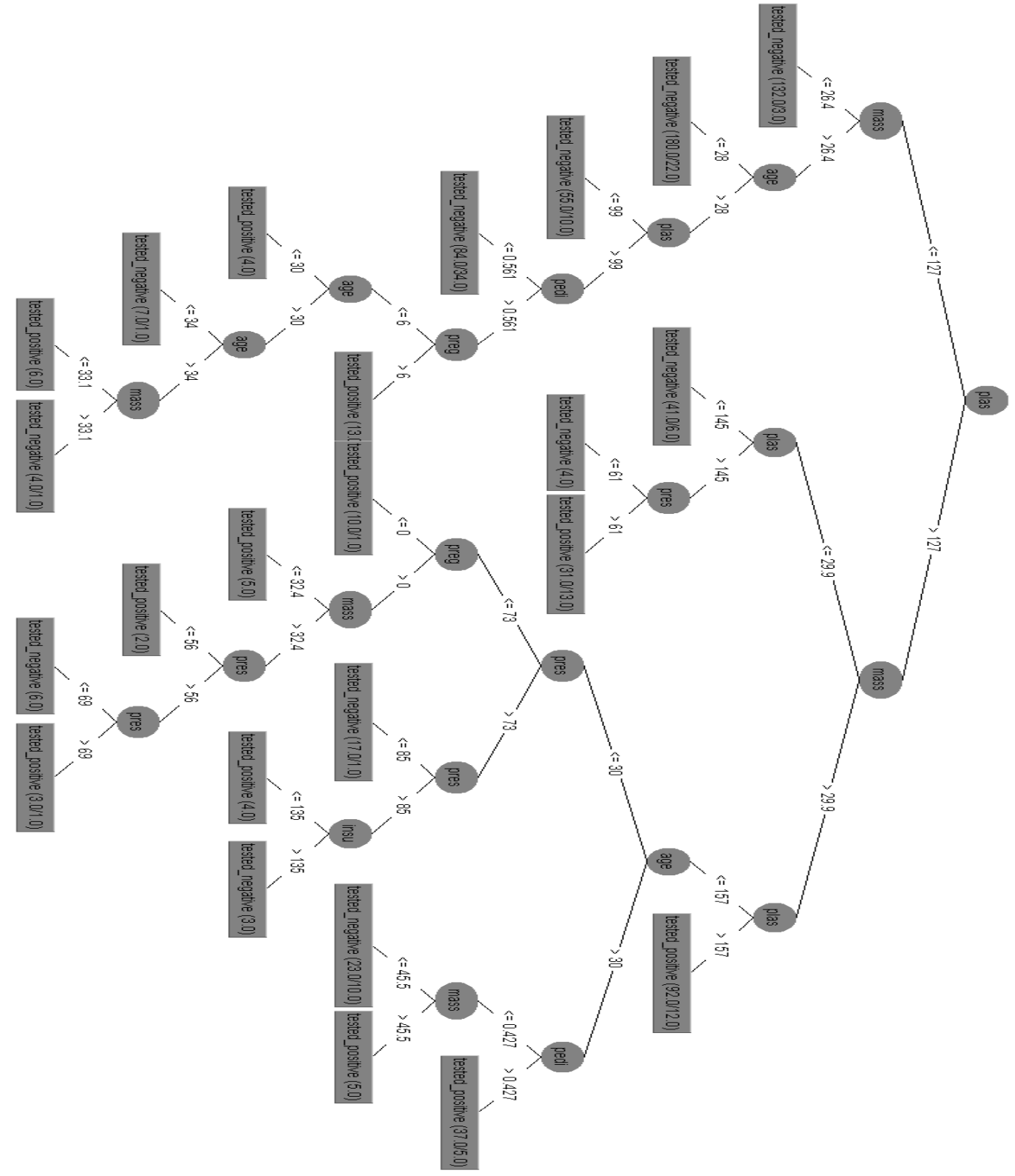


Figure – 4.2: Tree view of test environment-2.

III. Test Environment-3 :

g. Training data: New Data Set.

h. Test data: Raw Data Set.

i. Algorithm: C4.5

Here, new data set is using for training and raw data set as testing purpose. C4.5 is using as the algorithm. Table 4.7 is representing the detailed accuracy, table 4.8 is representing the summary of this test and table 4.9 is representing the confusing matrix by test environment-3. From the table 4.1.8 we can see that our technique successfully classify 647 instances which leads the 84.375 % accuracy.

Table – 4.7: Detailed accuracy of the new dataset for test environment-3.

			Weighted Avg.
Class	**Tested_Negative**	**Tested_Positive**	**---**
TP Rate	0.936	0.668	0.842
FP Rate	0.332	0.064	0.239
Precision	0.840	0.848	0.843
Recall	0.936	0.668	0.842
F-Measure	0.886	0.747	0.837
MCC	0.645	0.645	0.645
ROC Area	0.893	0.893	0.893
PRC Area	0.991	0.818	0.884

Table – 4.8: Summary of the new dataset for test environment-3.

	New Data
Correctly Classified Instances	647 (**84.375 %)**
Incorrectly Classified Instances	121 (15.625 %)
Kappa statistic	0.6353
Mean absolute error	0.2309
Root mean squared error	0.34
Relative absolute error	50.7966%
Root relative squared error	71.3321%
Total Number of Instances	768

Table – 4.9: Confusion matrix of the new dataset for test environment-3.

	New Data Set	
	Predicted: No	**Predicted: Yes**
Actual: No Tested_Negative	468	32
Actual: Yes Tested_Positive	89	179

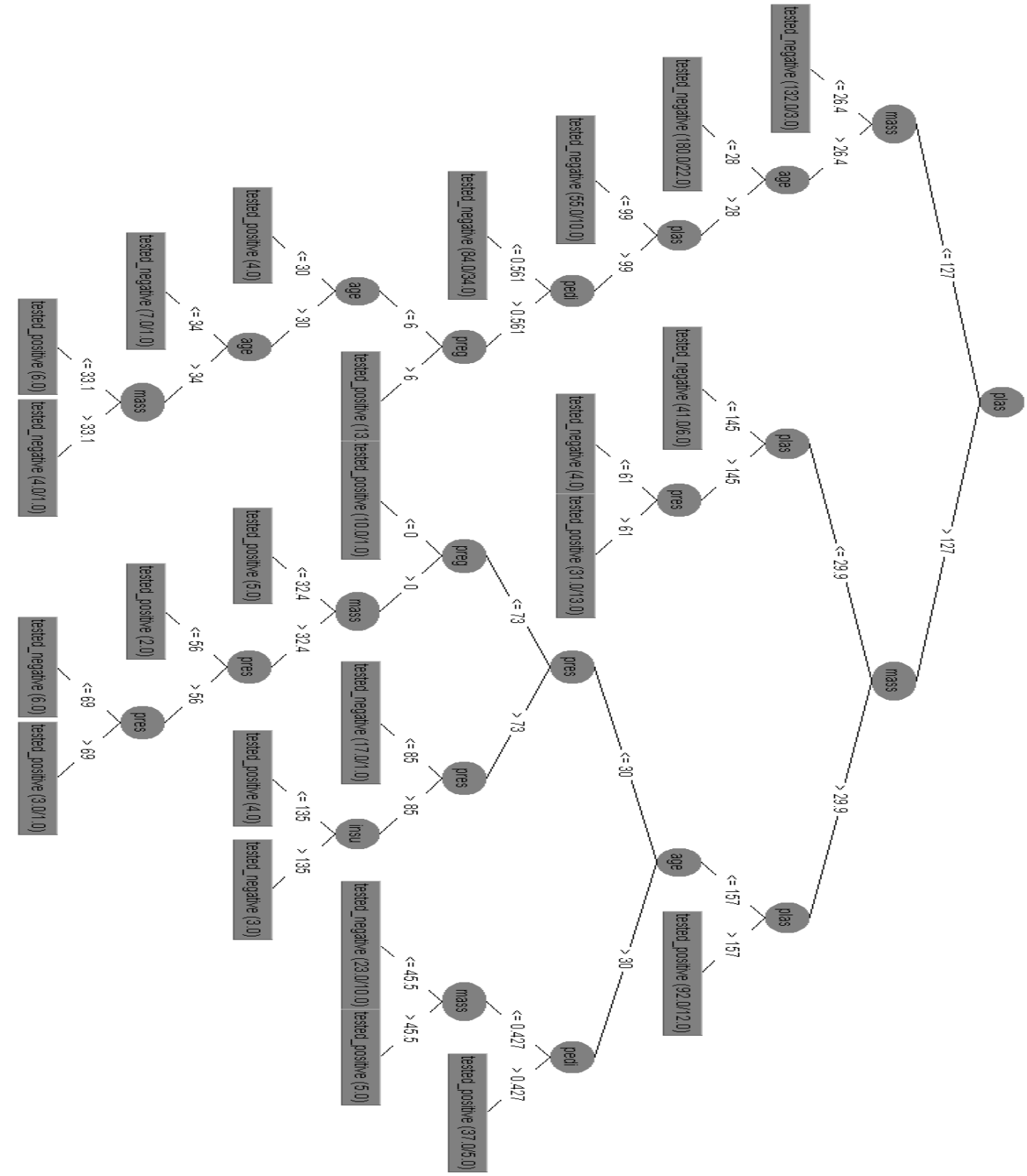


Figure – 4.3: Tree view of test environment-3.

IV. Test Environment-4 :

j. Training data: New Data Set.

k. Test data: Raw Data Set.

l. Algorithm: KNN. K=5.

Here, new data set is using for training and raw data set as testing purpose. KNN is using as the algorithm with the value of K=5. Table 4.1.10 is representing the detailed accuracy, table 4.1.11 is representing the summary of this test and table 4.1.12 is representing the confusing matrix by test environment-4.

Table – 4.10: Detailed accuracy of the new dataset for test environment-4.

			Weighted Avg.
Class	**Tested_Negative**	**Tested_Positive**	**---**
TP Rate	0.908	0.623	0.809
FP Rate	0.377	0.092	0.277
Precision	0.818	0.784	0.806
Recall	0.908	0.623	0.809
F-Measure	0.861	0.694	0.803
MCC	0.565	0.565	0.565
ROC Area	0.891	0.891	0.891
PRC Area	0.924	0.772	0.871

Table – 4.11: Summary of the new dataset for test environment-4.

	New Data
Correctly Classified Instances	621 (**80.8594 %)**
Incorrectly Classified Instances	147 (19.1406 %)
Kappa statistic	0.5577
Mean absolute error	0.2546
Root mean squared error	0.3554
Relative absolute error	56.0087 %
Root relative squared error	74.5617 %
Total Number of Instances	768

Table – 4.12: Confusion matrix of the new dataset for test environment-4.

	New Data Set	
	Predicted: No	**Predicted: Yes**
Actual: No Tested_Negative	454	46
Actual: Yes Tested_Positive	101	167

V. Test Environment-4 :

m. Training data: New Data Set.

n. Cross-validation with 10 folds.

o. Algorithm: KNN. K=5.

Here, new data set is using for training purpose and cross-validation with 10 folds. KNN is using as the algorithm with the value of K=5. Table 4.1.13 is representing the detailed accuracy, table 4.1.14 is representing the summary of this test and table 4.1.15 is representing the confusing matrix by test environment-5.

Table – 4.13: Detailed accuracy of the new dataset for test environment-5.

			Weighted Avg.
Class	**Tested_Negative**	**Tested_Positive**	**---**
TP Rate	0.848	0.526	0.736
FP Rate	0.474	0.152	0.326
Precision	0.770	0.650	0.728
Recall	0.848	0.526	0.736
F-Measure	0.807	0.581	0.728
MCC	0.396	0.396	0.396
ROC Area	0.781	0.781	0.781
PRC Area	0.844	0.627	0.768

Table – 4.14: Summary of the new dataset for test environment-5.

	New Data
Correctly Classified Instances	565 (**73.5677 %)**
Incorrectly Classified Instances	203 (26.4323 %)
Kappa statistic	0.3914
Mean absolute error	0.3092
Root mean squared error	0.4248
Relative absolute error	68.0353 %
Root relative squared error	89.1192 %
Total Number of Instances	768

Table – 4.15: Confusion matrix of the new dataset for test environment-5.

	New Data Set	
	Predicted: No	**Predicted: Yes**
Actual: No Tested_Negative	424	76
Actual: Yes Tested_Positive	127	141

4.2 PERFORMANCE COMPARISON

From this table, we can see that there is a huge number of missing and abnormal values in the “Pima Indiana Diabetic” data set. We apply our technique to finding those values to make this data set well organized. We take 40 mm Hg (figure-4.4) as the lower limit of the diastolic blood pressure [2]. So below 40 mm Hg we get thirty-nine (39) instances those need to be changed and we change them.

In the research, we use several test environments to test our framework and below every table is showing the difference between our result and raw data set result. Each test environment we attach one summary table, one confusion matrix table, and one figure to show the better view of accuracy.

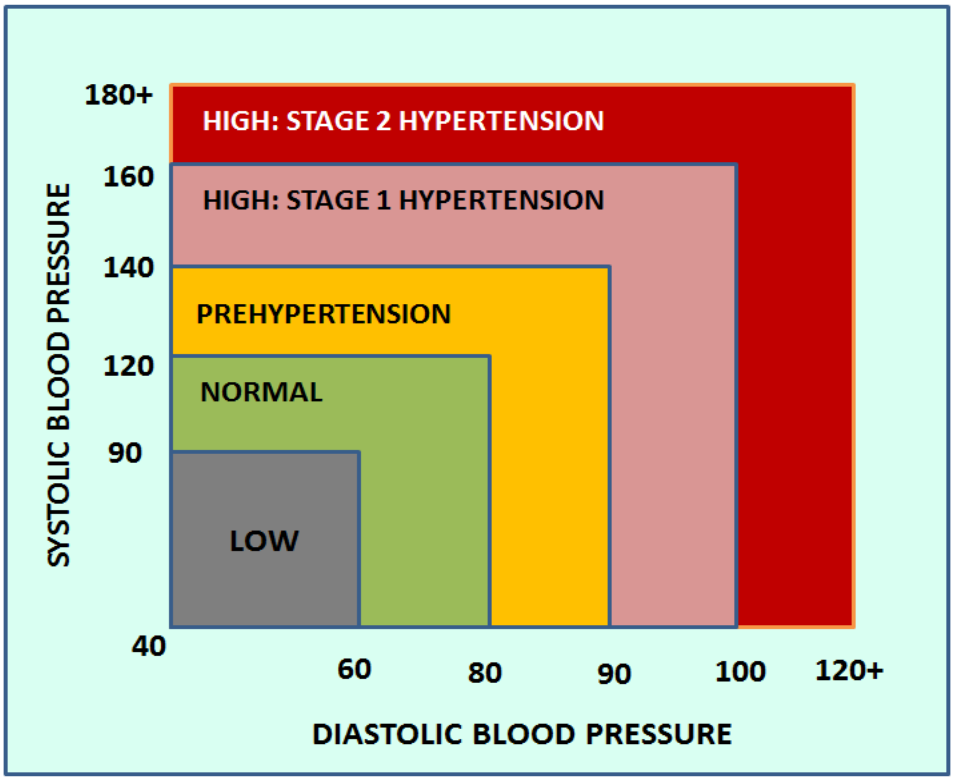


Figure - 4.4: Different blood pressure level.

Table - 4.16: Comparison of a single attributes (Diastolic blood pressure).

Raw data	New Data
Diastolic blood pressure (mm Hg)	**Diastolic blood pressure (mm Hg)**
Missing: 39 (5%) **Distinct: 43** Unique: 6(1%) Minimum : 40 Maximum: 122 **Mean: 72.635**	**Missing: 0 (0%)** **Distinct: 46** Unique: 6(1%) Minimum : 40 Maximum: 122 **Mean: 72.557**

I. Test Environment-1 :

a. Training data: New Data Set and Raw Data Set.

b. Test data: New Data Set and Raw Data Set.

c. Algorithm: C4.5

In the test environment-1, we use new data set as training and testing purpose and also we use raw data set as training and testing purpose. Here we use C4.5 as the testing algorithm.

From the below table we can see that we get higher accuracy by using our technique. Our accuracy is 84.375% but raw data set accuracy is 82.4219%. If we look at the relative absolute error then we can see that our error is less than the raw data set.

Table – 4.17: Summary of result for both dataset by test environment-1.

	Raw Data	**New Data**
Correctly Classified Instances	633 (**82.4219 %)**	648 (**84.375 %)**
Incorrectly Classified Instances	135 (17.5781 %)	120 (15.625 %)
Kappa statistic	0.5863	0.6386
Mean absolute error	0.2504	0.2288
Root mean squared error	0.3534	0.3382
Relative absolute error	55.0837%	50.3421%
Root relative squared error	74.1354%	70.9614%
Total Number of Instances	768	768

Table – 4.18: Confusion Matrix of both dataset by test environment-1.

	Raw Data Set		**New Data Set**	
	Predicted: No	**Predicted: Yes**	**Predicted: No**	**Predicted: Yes**
Actual: No Tested_Negative	470	30	468	32
Actual: Yes Tested_Positive	105	163	88	180

Below figure is showing the accuracy by graph view. We can see that the accuracy of new data set is higher than the raw data set.

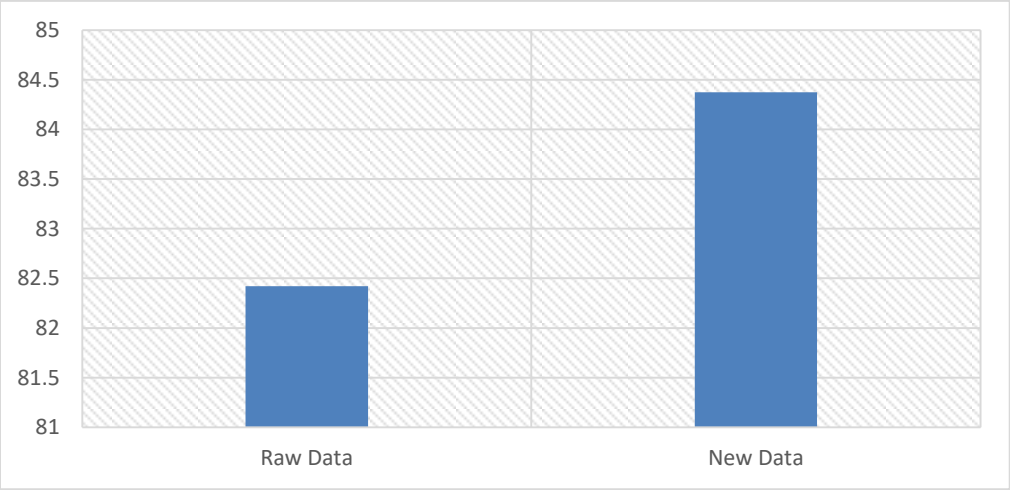


Figure - 4.5: Accuracy graph for test environment-1. The higher is better.

II. Test Environment-2:

d. Training data: New Data Set and Raw Data Set respectively.

e. Cross-validation with 10 folds.

f. Algorithm: C4.5

In the test environment-2, we use new data set as training and cross-validation with 10 folds and we also use raw data set as training and cross-validation with 10 folds. Here we use C4.5 as the testing algorithm.

From the below table we can see that we get higher accuracy by using our technique. Our accuracy is 73.4375% but raw data set accuracy is 73.1771%. If we look at the relative absolute error then we can see that our error is less than the raw data set.

Table - 4.19: Summary of result for both dataset by test environment-2.

	Raw Data	**New Data**
Correctly Classified Instances	562 (**73.1771 %)**	564 (**73.4375 %)**
Incorrectly Classified Instances	206 (**26.8229** %)	204 (**26.5625** %)
Kappa statistic	0.4056	0.4093
Mean absolute error	0.3182	0.3161
Root mean squared error	0.4479	0.4477
Relative absolute error	**70.0146**%	**69.541**%
Root relative squared error	93.9623%	93.9366%
Total Number of Instances	768	768

Table – 4.20: Confusion Matrix of both dataset by test environment-2.

	Raw Data Set		**New Data Set**	
	Predicted: No	**Predicted: Yes**	**Predicted: No**	**Predicted: Yes**
Actual: No Tested_Negative	401	99	404	96
Actual: Yes Tested_Positive	107	161	108	160

Below figure is showing the accuracy by graph view. We can see that the accuracy of new data set is higher than the raw data set.

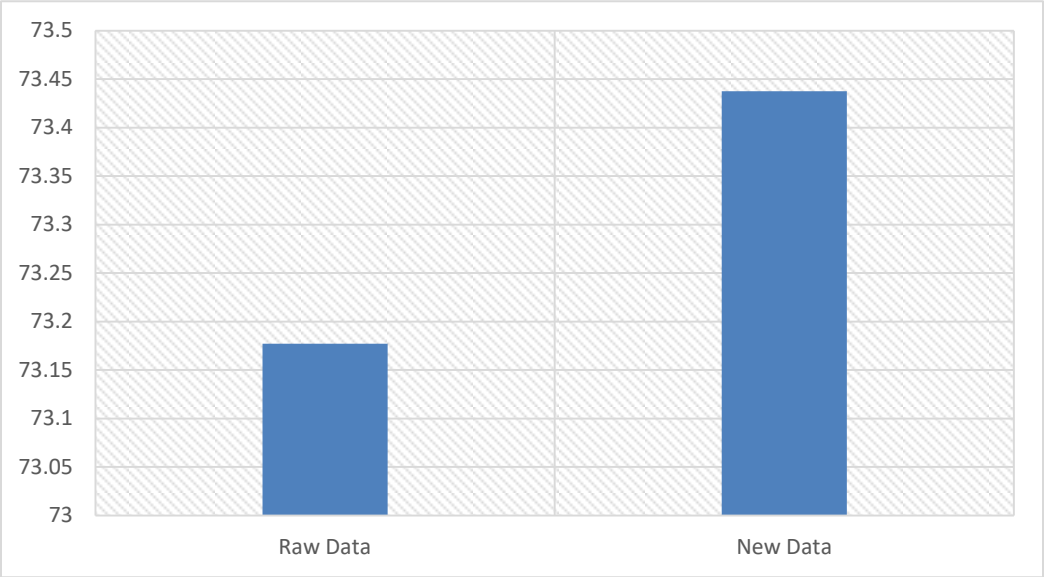


Figure-4.6: Accuracy graph by Test Environment-2. The higher is better.

III. Test Environment-3 :

g. Training data: New Data Set and Raw Data Set respectively.

h. Test data: Raw Data Set and New Data Set respectively.

i. Algorithm: C4.5

In the test environment-3, we use new data set as training and raw data set testing purpose and also we use raw data set as training and new data set as testing purpose. Here we use C4.5 as the testing algorithm.

From the below table we can see that we get higher accuracy by using our technique. Our accuracy is 84.2448% but raw data set accuracy is 82.4219%. If we look at the relative absolute error then we can see that our error is less than the raw data set.

Table - 4.21: Summary of result for both dataset by test environment-3.

	Raw Data	**New Data**
Correctly Classified Instances	633 (**82.4219 %)**	647 (**84.2448 %)**
Incorrectly Classified Instances	135 (17.5781 %)	121 (15.7552 %)
Kappa statistic	0.5863	0.6353
Mean absolute error	0.2506	0.2309
Root mean squared error	0.3539	0.34
Relative absolute error	55.1443 %	50.7966 %
Root relative squared error	74.2407 %	71.3321 %
Total Number of Instances	768	768

Table – 4.22: Confusion Matrix of both dataset by test environment-3.

	Raw Data Set		**New Data Set**	
	Predicted: No	**Predicted: Yes**	**Predicted: No**	**Predicted: Yes**
Actual: No Tested_Negative	470	30	468	32
Actual: Yes Tested_Positive	105	163	89	179

Below figure is showing the accuracy by graph view. We can see that the accuracy of new data set is higher than the raw data set.

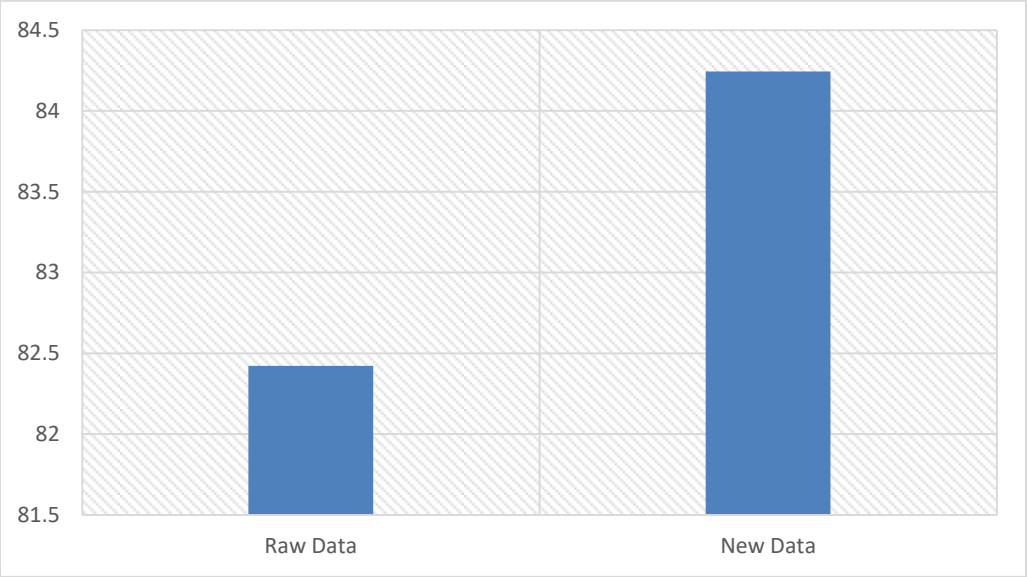


Figure – 4.7: Accuracy graph for Test Environment-3. The higher is better.

IV. Test Environment-4 :

- **j.** Training data: New Data Set and Raw Data Set respectively.
- **k.** Test data: New Data Set and Raw Data Set respectively.
- **l.** Algorithm: KNN. K=5.

In the test environment-4, we use new data set as training and testing purpose and also we use raw data set as training and testing purpose. Here we use KNN as the testing algorithm with the value of K=5.

From the below table we can see that we get higher accuracy by using our technique. Our accuracy is 81.6406% but raw data set accuracy is 80.9896%. If we look at the relative absolute error then we can see that our error is less than the raw data set.

Table - 4.23: Summary of result for both dataset by test environment-4.

	Raw Data	**New Data**
Correctly Classified Instances	622 (**80.9896 %)**	627 (**81.6406 %)**
Incorrectly Classified Instances	146 (19.0104 %)	141 (18.3594 %)
Kappa statistic	0.5603	0.5796
Mean absolute error	0.2553	0.2486
Root mean squared error	0.3568	0.3488
Relative absolute error	56.1806 %	54.6916 %
Root relative squared error	74.853 %	73.1812 %
Total Number of Instances	768	768

Table – 4.24: Confusion Matrix of both dataset by test environment-4.

	Raw Data Set		**New Data Set**	
	Predicted: No	**Predicted: Yes**	**Predicted: No**	**Predicted: Yes**
Actual: No Tested_Negative	455	45	452	48
Actual: Yes Tested_Positive	101	167	93	175

Below figure is showing the accuracy by graph view. We can see that the accuracy of new data set is higher than the raw data set.

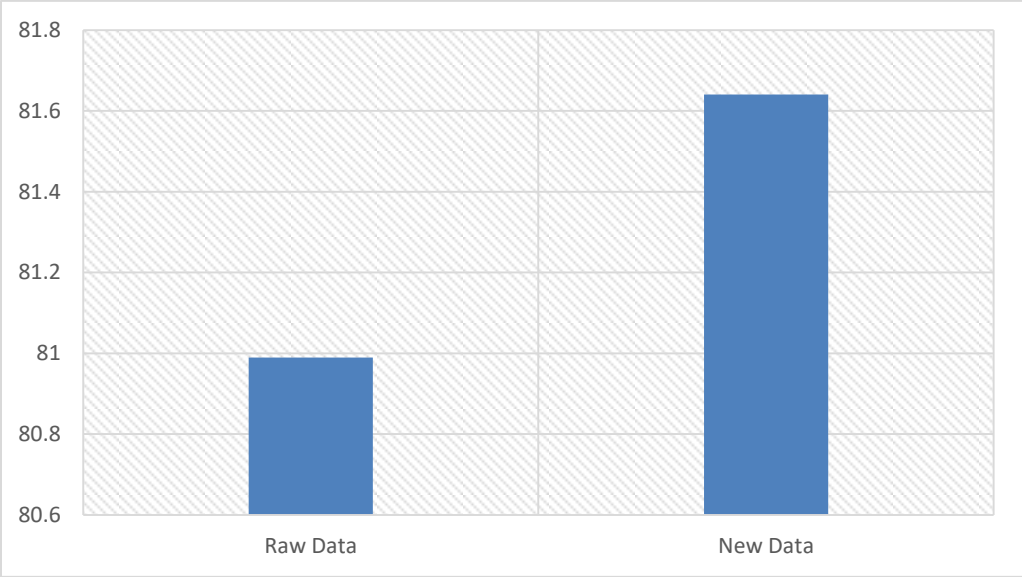


Figure – 4.8: Accuracy graph for Test Environment-4. The higher is better.

V. Test Environment-5 :

m. Training data: New Data Set and Raw Data Set respectively.

n. Cross-validation with 10 folds.

o. Algorithm: KNN. K=5.

In the test environment-5 we use new data set as training and cross-validation with 10 folds and we also use raw data set as training and cross-validation with 10 folds. Here we use KNN as the testing algorithm with the value of K=5.

From the below table we can see that we get higher accuracy by using our technique. Our accuracy is 73.5677% but raw data set accuracy is 72.7865%. If we look at the relative absolute error then we can see that our error is less than the raw data set.

Table - 4.25: Summary of result for both dataset by test environment-5.

	Raw Data	**New Data**
Correctly Classified Instances	559 (**72.7865 %)**	565 **(73.5677 %)**
Incorrectly Classified Instances	209 (27.2135 %)	203 (26.4323 %)
Kappa statistic	0.3642	0.3914
Mean absolute error	0.3165	0.3092
Root mean squared error	0.431	0.4248
Relative absolute error	69.6387 %	68.0353 %
Root relative squared error	90.4211 %	89.1192 %
Total Number of Instances	768	768

Table – 4.26: Confusion Matrix of both dataset by test environment-5.

	Raw Data Set		**New Data Set**	
	Predicted: No	**Predicted: Yes**	**Predicted: No**	**Predicted: Yes**
Actual: No Tested_Negative	429	71	424	76
Actual: Yes Tested_Positive	138	130	127	141

Below figure is showing the accuracy by graph view. We can see that the accuracy of new data set is higher than the raw data set.

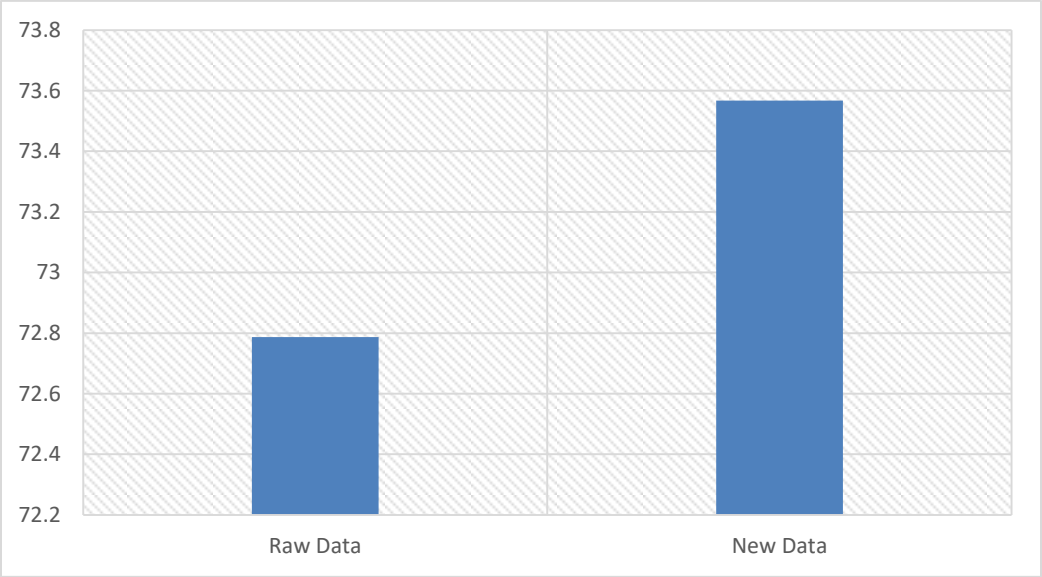
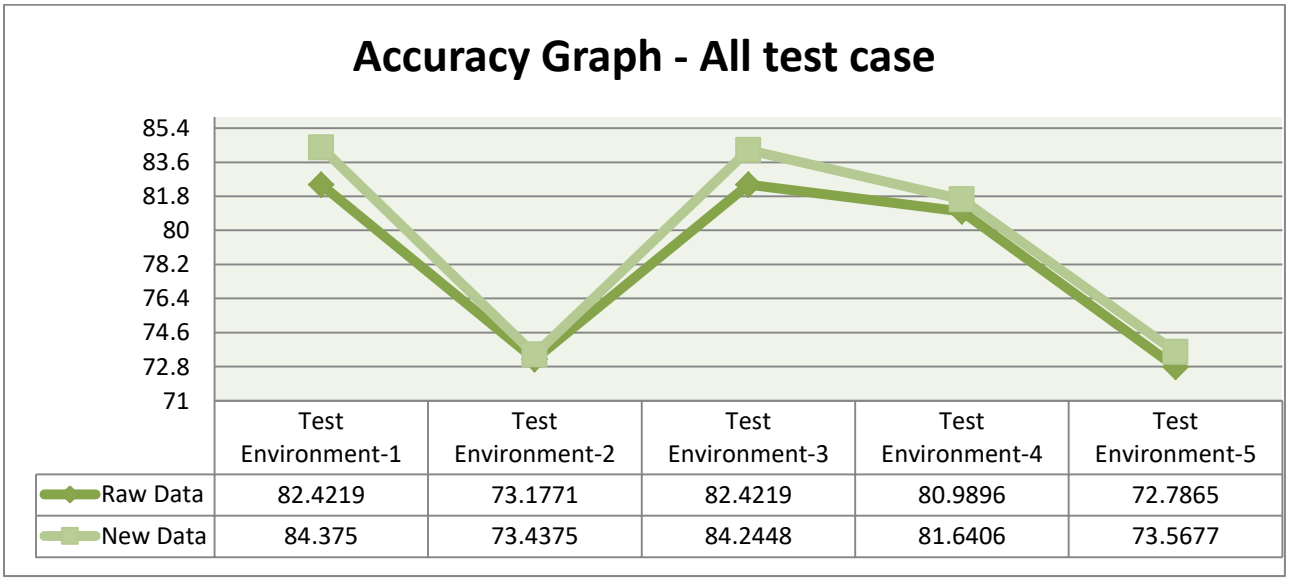


Figure- 4.9: Accuracy graph for Test Environment-5. The higher is better.

Overall differences in all the test environment in a graph is given below. From the below figure we can see that we get better accuracy in every test we perform.



	Test Environment-1	Test Environment-2	Test Environment-3	Test Environment-4	Test Environment-5
Raw Data	82.4219	73.1771	82.4219	80.9896	72.7865
New Data	84.375	73.4375	84.2448	81.6406	73.5677

Figure- 4.10: Accuracy graph by all test environments.

4.3 SUMMARY

In this research, we have also a web application where we demonstrate that how our proposed system will work in real life. Here we include hospitals, users and other entities at the same place. If someone uses this application than they will get all the modern facilities according to their need.

A member of the hospital will fill up this form (Appendix A1 & Appendix A2) in the pathology laboratory and when they will click on submit button this record will be saved in the central database which is control by the government. A user also can send data from their mobile phone to this central system.

They have access to edit or update the record which was submitted by them (Hospital-1). They can also view the record that only submitted by them (Hospital-1). When they will click on the edit button (Appendix A3) a new form (Appendix A4) will arrive with edit functionality and after complete the process they have to click on submit button.

Another part of this system is patient data availability to the law enforcement agencies. When law enforcement agency will enter a unique id (National ID or any ID) in the search box (Appendix A5) they will see the result (Appendix A6) of that patient. They cannot edit it; they will get only view access. Without government permission, no one will get to view any information from this system.

Insurance companies are very interested in our health data. If someone goes to the insurance company and ask for insurance than they have to fill up a form and then the company will verify it from this system. The insurance company also can view this information only if the users want to send their data to the company.

CHAPTER 5

CONCLUSION

5.1 OVERALL CONCLUSION

In this research, we develop an intelligent system for predicting the diseases like diabetes, heart, cancer so on. We also develop a system that can be integrated with any existing machine learning algorithms for classification task. Finally, we introduce IoT with all the above topics discussed. Here everything is executed and implemented in real life scenario. In this experiment, we perform several test environments and we get the better result in every case for example, in the test environments-1 we get 84.375% accuracy where raw dataset achieve 82.4219% accuracy and we also able to reduce the relative absolute error from 55.0831% to 50.3421%. It means 4.7416% error reduction from test environment-1. Similarly, we have succeeded in every test environment. This system will give the benefits to doctors, clinical company, medical students, people who work with IoT.

5.2 FUTURE WORKS

Every research has carried further exploration and more research and this research is no exception. In this research, we use our own technique so this is possible to use another technique to get a better accuracy but our technique is given better accuracy among all the studied papers. So in future, we will apply more technique to increase the accuracy. We don’t have available IoT devices and those are very costly so when they will be available and will be cheap we will make more users friendly system. In this research, we haven’t implement bio data security, so this could be a good study option for future research.

REFERENCES

[1] Begum, Syeda Faiza Unnisa. Begum, Imtiyazunnisa. Smart Health Care Solutions Using IOT. *International journal and magazine of engineering, technology, management and research*, vol. 4, no.3, 2017

[2] "STAY WELL WORLD | Dignity of the human person", *STAY WELL WORLD | Dignity of the human person*, 2017. [Online]. Available: https://www.staywellworld.org/single-post/2017/06/03/LOW-BLOOD-PRESSURE-HYPOTENSION. [Accessed: 02- Aug- 2017].

[3] M. Haghi, K. Thurow and R. Stoll, "Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices", *Healthcare Informatics Research*, vol. 23, no. 1, p. 4, 2017.

[4] "Hicon smartwristband with social network icons", *Indiegogo*, 2017. [Online]. Available: https://www.indiegogo.com/projects/hicon-smartwristband-with-social-network-icons#/. [Accessed: 02- Aug- 2017].

[5] b. Shepard, "News - Low diastolic blood pressure linked to higher risk of heart failure", *Uab.edu*, 2017. [Online]. Available: https://www.uab.edu/news/latest/item/1681-low-diastolic-blood-pressure-linked-to-higher-risk-of-heart-failure. [Accessed: 02- Jul- 2017].

[6]"Blood Pressure : Q. What is low blood pressure?", *Bloodpressureuk.org*, 2017. [Online]. Available: http://www.bloodpressureuk.org/microsites/u40/Home/facts/Whatislow. [Accessed: 02- Aug- 2017].

[7] R. Pushkaraj Bhandari, P. Sapna Yadav, A. Shyam Mote, and P. Devika Rankhambe. "A survey on predictive system for medical diagnosis with expertise analysis". *International journal of innovative research in computer and communication engineering*, vol 4, 2016.

[8] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining", *Indian Journal of Science and Technology*, vol. 9, no. 43, 2016.

[9] K.U. Sreekanth., K.P. Nitha., A Study on Health Care in Internet of Things. *International Journal on Recent and Innovation Trends in Computing and Communication*, vol.4, no.2, 2016.

[10] D. Dziak, B. Jachimczyk and W. Kulesza, "IoT-Based Information System for Healthcare Application: Design Methodology Approach", *Applied Sciences*, vol. 7, no. 6, p. 596, 2017.

[11] L. Yehia, A. Khedr and A. Darwish, "Hybrid Security Techniques for Internet of Things Healthcare Applications", *Advances in Internet of Things*, vol. 05, no. 03, pp. 21-25, 2015.

[12] G. Huang, K. Huang, T. Lee and J. Weng, "An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients", *BMC Bioinformatics*, vol. 16, no. 1, p. S5, 2015.

[13] A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 01-14, 2015.

[14] S. Riazul Islam, Daehan Kwak, M. Humaun Kabir, M. Hossain and Kyung-Sup Kwak, "The Internet of Things for Health Care: A Comprehensive Survey", *IEEE Access*, vol. 3, pp. 678-708, 2015.

[15] P. Patidar and A. Tiwari, "Handling Missing Value in Decision Tree Algorithm", *International Journal of Computer Applications*, vol. 70, no. 13, pp. 31-36, 2013.

[16]A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime", *Procedia Computer Science*, vol. 85, pp. 662-668, 2016.

[17] R.K. Gupta and P. Singh, "A New Way to Design and Implementation of Hybrid Crypto System for Security of the Information in Public Network. *International Journal of Emerging Technology and Advanced Engineering*, vol.3, 2250-2459, 2013.

[18] Yu, L.L., Wang, Z.J. and Wang, W.F. (2012). " The Application of Hybrid Encryption Algorithm in Software Security". *2012 4th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp.762-765, 2012.

[19] K. Raja, G. Tholkappia Arasu, Chitra. S. Nair, "Imputation Framework for Missing Values". International Journal of Computer Trends and Technology, vol.3, no.2, pp.215-219, 2012.

[20]P. Anooj, "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules", *Open Computer Science*, vol. 1, no. 4, 2011.

[21] D. Senthil Kumar, G. Sathyadevi, and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining". *International Journal of Computer Science Issues*, vol.8, no.3, 2011.

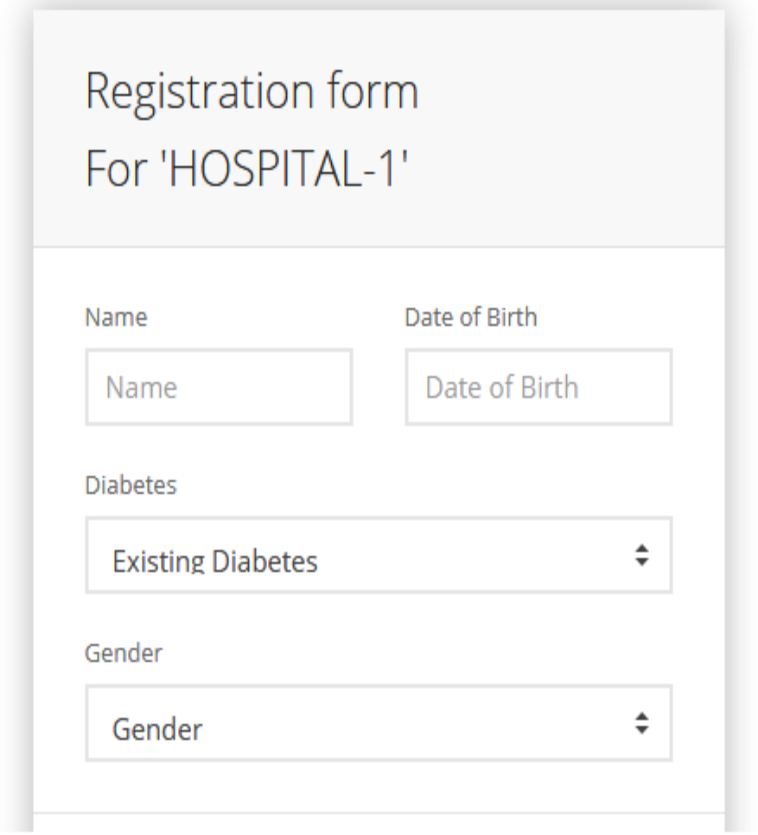
[22] M. Saar-Tsechansky, F. Provost, "Handling missing values when applying classification models". *Journal of machine learning research*, 2007.

[23] B. Mehala., P. Ranjit Jeba Thangaiah., and K. Vivekanandan. "Selecting Scalable Algorithms to Deal With Missing Values". *International Journal of Recent Trends in Engineering*, vol.1, no.2, 2009.

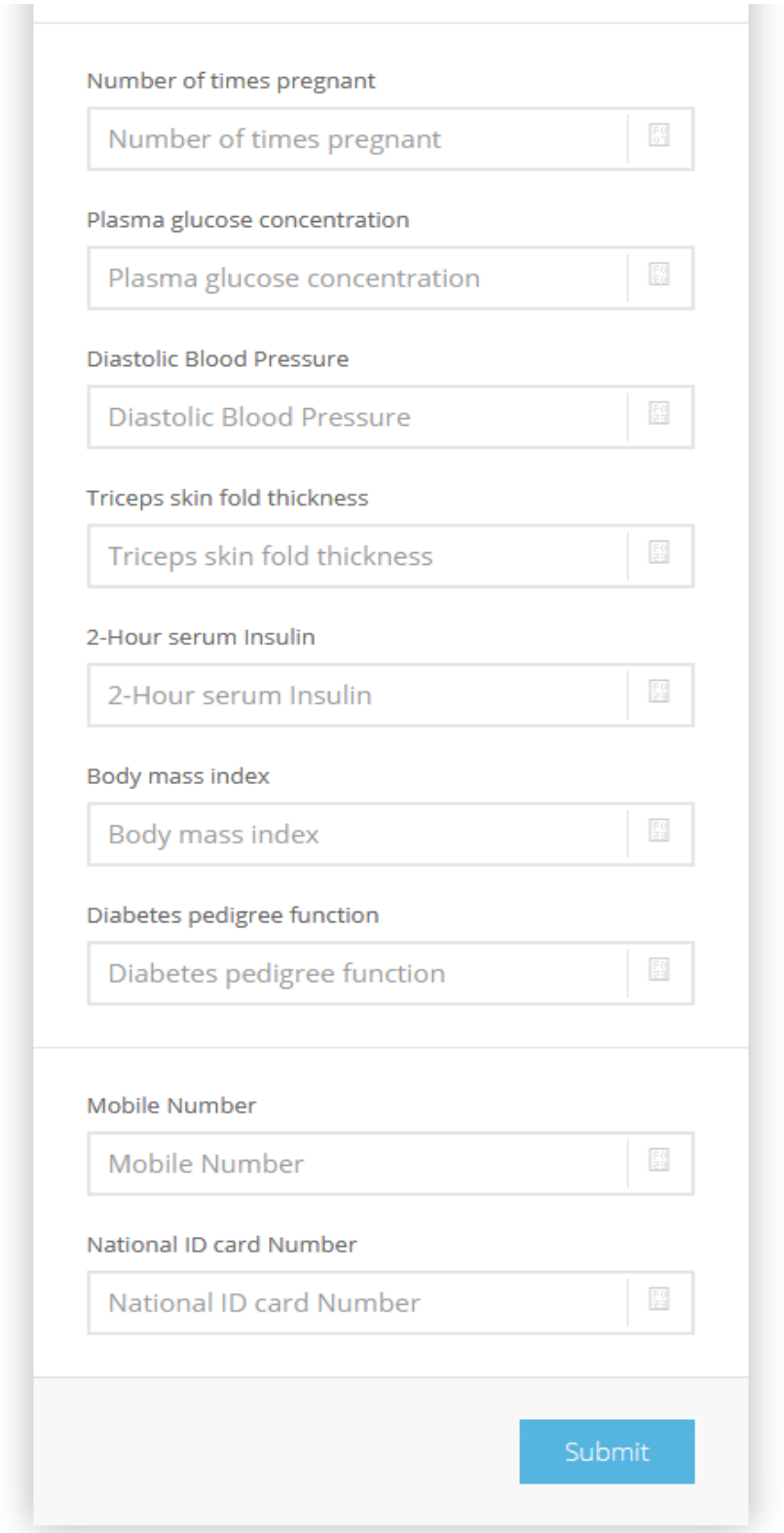
[24]U. DBD, "Confusion Matrix", *Www2.cs.uregina.ca*, 2017. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html. [Accessed: 02- Aug- 2017].

[25] "UCI Machine Learning Repository: Pima Indians Diabetes Data Set", *Archive.ics.uci.edu*, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. [Accessed: 02- Aug- 2017].

Appendix A1: Registration Form (Part A).



Appendix A2: Registration Form (Part B).



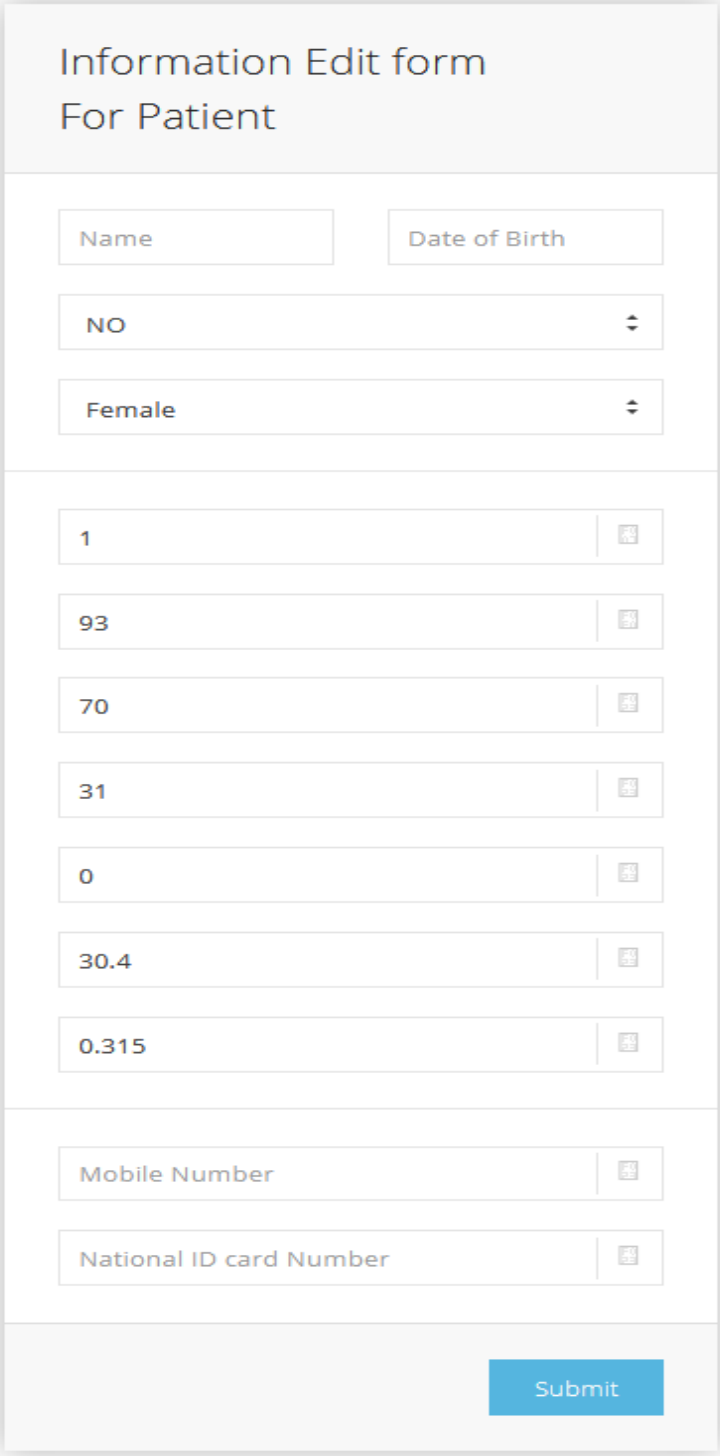
Appendix A3: Admin Panel.

Admin Panel

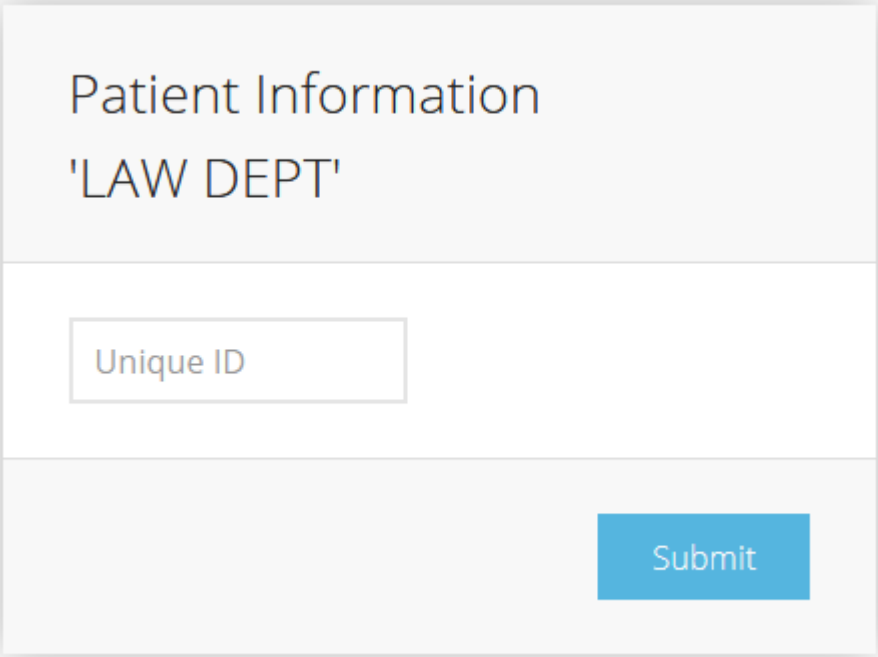
Index Users Total Record View Download All data Download Raw data Import All data Import Raw data
Logout

Diastolic Blood Pressure	Trice	Insulien	dp function	age	Dabetic or Not	Action
70	31	0	0.315	23	0	Edit | Delete
60	0	0	0.349	47	1	Edit | Delete
72	23	112	0.245	30	0	Edit | Delete
70	27	0	0.34	27	0	Edit | Delete
76	48	180	0.171	63	0	Edit | Delete
62	0	0	0.142	33	0	Edit | Delete
74	31	0	0.403	43	1	Edit | Delete
58	26	16	0.766	22	0	Edit | Delete
92	0	0	0.278	66	1	Edit | Delete
76	0	0	0.197	26	0	Edit | Delete
72	0	0	0.258	52	1	Edit | Delete
90	41	0	0.391	39	0	Edit | Delete

Appendix A4: Information Edit Form.



Appendix A5: Inquiry Form.



Appendix A6: Patient Information View Page.