KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

BY

Monira Akter Laboni

ID: 151-15-5044

AND

Hajera Akter

ID: 151-15-4819

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering (CSE)

Supervised By

Syed Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOV 2018

APPROVAL

This Project titled **"Kidney Disease Prediction Using Machine Learning"**, submitted by Monira Akter Laboni and Hajera Akter to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (CSE) and approved as to its style and contents.

BOARD OF EXAMINERS

_____

Dr. Syed Akther Hossain — **Chairman**
Professor and Head
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor and Associate Head
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science & Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science & Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Syed Akhter Hossain, Professor and Head,** Department of Computer Science and Engineering (CSE), Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Syed Akhter Hossain
Professor and Head
Department of CSE
Daffodil International University

Submitted by:

_____

Monira Akter Laboni
ID: 151-15-5044
Department of CSE
Daffodil International University

_____

Hajera Akter
ID: 151-15-4819
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we want to express our heartiest thanks and gratefulness to the Almighty Allah for His divine blessing makes us possible to complete the final year project successfully.

We are really grateful and wish our profound indebtedness to **Dr. Syed Akhter Hossain, Professor and Head,** Department of CSE of Daffodil International University, Dhaka. Deep Knowledge and keen interest of our supervisor in the active learning model design influenced to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would also like to express our heartiest gratitude to other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Machine learning has earned a remarkable position in healthcare sector because of its capability to enhance the disease prediction in healthcare sector. Artificial intelligence and Machine learning techniques are being used in healthcare sector. Nowadays, kidney disease (KD) is becoming a major public health problem worldwide. It is increasing day by day because of not maintaining proper food habits, drinking less amount of water and lack of health consciousness. For that reason, there should have one or more approaches that can effectively keep tracking and monitoring people's kidney and health condition in an application view. Here, we have proposed an approach for real time kidney disease prediction, monitoring and application (KDPMA). Our aim is to develop an optimized and efficient machine learning (ML) application that can effectively recognize and predict the condition of chronic kidney disease. In this work, ten most important machine learning classification techniques were considered for predicting chronic kidney disease. In this process, the data has been divided into two sections. In one section train dataset got trained and another section got evaluated by test dataset. The analysis results show that Decision Tree Classifier and Gaussian Naïve Bayes achieved highest performance than the other classifiers, obtaining the F1 measure of 1.0. In future, we will make an application based on the best output results classifier technique to predict Kidney Disease.

©Daffodil International University

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-3**
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Report Layout	3
CHAPTER 2: BACKGROUND	**4-8**
2.1 Introduction	4
2.2 Related Works	4
2.3 Research Summary	4
2.3 Scope of the Problem	8
2.4 Challenges	8
CHAPTER 3: RESEARCH METHODOLOGY	**9-14**
3.1 Introduction	9
3.2 Research Subject and Instrumentation	9

©Daffodil International University

3.3 Data Collection Procedure 9

3.4 Statistical Analysis 9

3.5 Implementation Requirements 10-13

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **14-23**

4.1 Introduction 14

4.2 Experimental Results 15

4.3 Descriptive Analysis 16

4.4 Summary 17-23

CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **24-24**

5.1 Summary of the Study 24

5.2 Conclusions 24

5.3 Recommendations 24

5.4 Implication for Further Study 24

REFERENCES 25

APPENDIX 26

PLAGIARISM REPORT SCREENSHOT 27

LIST OF FIGURES

FIGURES **PAGE NO**

Figure 2.2.1: Features of others Research Work 7

Figure 2.2.2: Classification procedure of others research 8

Figure 3.4.1: Flowchart for KD Analysis and Prediction 12

Figure 4.1.1: Dataset Chart Ratio 14

Figure 4.1.2: Accuracy, Sensitivity and Specificity 15

Figure 4.2.1: Accuracy of K-Neighbor Classifier 17

Figure 4.2.2: Accuracy of Support Vector Classifier 18

Figure 4.2.3: Accuracy of Decision Tree Classifier 18

Figure 4.2.4: Accuracy of Random Forest Classifier 19

Figure 4.2.5: Accuracy of Ada Boost Classifier 19

Figure 4.2.6: Accuracy of Gradient Boosting Classifier 20

Figure 4.2.7: Accuracy of Gaussian Naïve Bayes Classifier 20

Figure 4.2.8: Accuracy of Linear Discriminant Analysis Classifier 20

Figure 4.2.9: Accuracy of Logistic Regression Classifier 20

Figure 4.2.10: Accuracy of Artificial Neural Network Classifier 20

Figure 4.3.1: Correlation Between Different Predictors 20

Figure 4.3.2: Classifier Accuracy 20

©Daffodil International University

CHAPTER 1

INTRODUCTION

1.1 Introduction

Kidney disease, also called kidney failure, describes the gradual loss of kidney function. Our kidneys filter wastes and excess fluids from our blood, which are then excreted in your urine. When kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in your body. In the early stages of kidney disease, you may have few signs or symptoms. Kidney disease (KD) may not become apparent until your kidney function is significantly impaired. Treatment for chronic kidney disease focuses on slowing the progression of the kidney damage, usually by controlling the underlying cause. Kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant.

On the other hand, classification is one of the most widely used methods of machine learning in healthcare organization. The classification technique predicts the target class for each data points. The classification methods such as Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes and Neural Network. Decision Tree is widely used by many researchers in healthcare field like skin diseases and chronic kidney disease etc. The K-Nearest Neighbor is used to analyze the relationship between cardiovascular disease, hypertension and the risk factors of various chronic diseases in order to construct an early warning system. Multilayer Neural Network is used for diagnosis of various chest related diseases such as Lung Cancer, Asthma, and Pneumonia etc.

Our aim is to predict the kidney disease, analyzing the data using the most popular ten classification techniques of machine learning that which techniques give us the maximum rate of accuracy to predict the disease. The ten classification techniques are K-Neighbors Classifier, Support Vector Classifier (SVC), Decision Tree Classifier (DT), Random Forest Classifier (RF), Ada Boost Classifier, Gradient Boosting Classifier, Gaussian Naïve Bayes, Linear Discriminant Analysis, Logistic Regression and Artificial Neural Network (ANN).

The objectives of our thesis are given below:

- To study how to classify or categorize disease data using some classifier algorithms
- To get a suitable technique that will be give the highest accuracy rate to predict the disease
- To visualize some analytical analysis of Kidney Disease classification classified by classifier algorithms
- To know about most popular classification algorithms of machine learning

1.2 Motivation

Kidney disease is rapidly growing disease in Bangladesh. It is recognized as a major public health problem in Bangladesh. There are various types of kidney diseases occurring in the world. The aim of this study is to predict the rate of being affected by chronic kidney disease. The study also showed the age distribution among the patients where majority was in between 46 years to 55 years old. Results of this study also demonstrated that the patient either have KD or do not have KD. The result of this study is expected to improve the awareness and consciousness to the people regarding the causes and consequences of KD, which ultimately will help to progress the consciousness and awareness among the mass people Bangladesh. Analyzing the entire data machine learning classification techniques will give the maximum rate of accuracy which will help to make an application based on that in general classification techniques.

Besides this, we see that today's world is so much focusing on recommendation system. Users expect everything that the better things will be recommended to them by the system. To make a system to be recommendation capable must have the ability to take decision by itself. To take decision by itself must need to have classified data.

All the above reasons made us interested to do such kind of research based work. Our work is firmly related to machine learning techniques and has some data mining procedures too.

1.3 Rationale of the Study

It is no doubt there are lots of works on Kidney Disease (KD) in English and these approaches or the processes are being used in many automated system as well as robotics system. But, these much classification techniques working on kidney disease data is very rare. To develop more automated application or make much more efficient of Machine Learning approaches in Kidney Disease, there has no alternative to work with the classified data. This made us to be interested to work with this classification.

In the present time, we see that there are so many researches on this kidney disease but all those are not that much distinct from each other whereas we are using ten different classification techniques to find out the best accuracy results which techniques give to make a further application in future.

1.4 Research Question

- Can we collect raw data of Kidney Disease?
- Can we pre-process the raw data to be used for the Machine Learning approaches?
- Can the Machine Learning process correctly detect or identify the category of the given Kidney Disease dataset?

1.5 Expected Output

Expected outcome of this research based project is to build an algorithm or making a complete efficient procedure that will categorize given Kidney Disease dataset with respect to the built model of trained dataset.

1.6 Report Layout

The report will be followed as following

Chapter 1 provides the summary of this research based project. Introductory discussion is the key term of this first chapter. Apart from, what motivated us to do such a research based project is explained well in this chapter to. The most important part of this chapter is the Rationale of the Study. Then, what are the research questions and what is the expected outcome is discussed in the last section of this chapter.

Chapter 2 covers the discussion on what already done in this domain before. Then the later section of this second chapter shows the scope arisen from their limitation of this field. And very last, the root obstacles or challenges of this research are explained.

Chapter 3 is nothing but the theoretical discussion on this research work. To discuss the theoretical part of the research, this chapter elaborates the statistical methods of this work. Besides, this chapter shows the procedural approaches of the Machine Learning classifier techniques. And in the last section of this chapter, to validate the model as well as to show the accuracy label of the classifier, confusion matrix analysis is being presented.

Chapter 4 is related with the outcome of the whole research and the project. Some experimental pictures are presents in this chapter to make realize the project.

Chapter 5 is based on conclusion topics of the project. This chapter is responsible to show the whole project report adhering to recommendation. The chapter is closed by showing the limitations of our works that can be the future scope of others who want to work in this field.

CHAPTER 2
BACKGROUND

2.1 Introduction

This chapter reflects the related works that already done by some researchers in the previous time in this field. Besides, giving a clear explanation of this, this chapter will show what the limitations of these works were and lastly, this chapter describes scope of our research as well as the challenges of it.

2.2 Related Works

It is the matter of sorrow that very few works in this field has accomplished by this time though in the present time, working on this field is increasing day by day. There are enough resources for Kidney Disease [1] as there has been done many works in this field.

Recently, not only in Kidney disease, but also in other diseases as like diabetes prediction [2], heart disease [3], and so on are being included on Disease Prediction using machine learning related works. There are being enriched with resources day by day after doing more research works on this field.

Some related works relates to our research work are given below with a short description.

Prediction of Kidney Disease Using Data Mining Techniques

Joining a machine learning calculation with conventional factual displaying, and furthermore planning a perplexing overview and changing the missing information are the principle commitment of this examination. The highlights they considered in planning a survey incorporate sex, age, race, smoking, nourishment security, Poverty Income Ratio, Body Mass Index, physical movement, liquor utilize, medicinal conditions and meds [4]. Forecast of four kinds of Kidney sicknesses in particular Nephritic Syndrome, Kidney illness, Acute Renal Failure and Glomerulonephritis. Administered characterization calculation Support Vector Machine (SVM) and Artificial Neural Network (ANN) is utilized to foresee the kidney ailment.

Exploratory outcomes demonstrate that ANN is best classifier Classification exactness for ANN is higher contrasted with SVM and the execution time for SVM is bring down contrasted with ANN.

ANN has better arrangement exactness [5]. Distinctive machine learning characterization calculation for analysis of unending kidney sickness is talked about. Different grouping procedures that have been utilized are: Decision Tree, Linear Discriminant classifier, Quadratic Discriminant classifier, Linear SVM, Quadratic SVM, Fine KNN, Medium KNN, Cosine KNN, Cubic KNN [6], Weighted KNN, Feed Forward Back Propagation Neural Network utilizing Gradient Descent and Feed Forward Back Propagation Neural [7]

An examination concentrate to uncover the significance of highlight/characteristic choice in classifier execution is depicted. Different order calculations: Sequential Minimal Optimization, Naïve Bayes and k-closest neighbor calculation (IBK) [8] classifiers were utilized to group KD patients with non-KD patients. WEKA was utilized as information mining apparatus. Wrapper subset property evaluator and best first scan technique were utilized for highlight determination. Results demonstrated that the execution of the classifiers enhanced in the wake of decreasing the quantity of highlights. IBK classifier played out the best contrasted with the others on a diminished dataset [9]. Anticipating of quality of Kidney Disease dependent on certain wellbeing parameters, for example, arbitrary blood glucose level l[10], serum creatinine level, pulse and others. Missing qualities in the dataset were ascribed with the normal estimation of the relating highlight section. Arrangement calculations were connected to precisely foresee the nearness kidney malady (KD) and bunching calculation were utilized to assemble the information dependent on the nearness of KD. Distinctive characterization calculations that were actualized are choice tree, strategic relapse, Ada lift and Support Vector Machine. Central part investigation was utilized to lessen the measurements and bunching calculation, for example, K-implies [11] and progressive grouping were connected. Exploratory outcomes demonstrated that the execution of SVM (with straight portion) was the best pursued by Random Forest Classifier, Ada support, Logistic relapse and Decision tree [12].

Test Data

The dataset used in this paper has been obtained from UCI source [13]. The dataset contains data of 400 samples from the southern part of India with their ages ranging between 2-90 years. There are in total twenty-four features, most of which are clinical in nature and the rest are physiological. Table 1 summarizes various parameters. As a part of data pre-processing, missing values and outliers are imputed with mean value of that feature for continuous data and attribute model value for categorical data. Nominal data are converted to numerical values. For example, Nominal values 'Normal' are labeled "1" and 'Abnormal' are labeled "0".

1	Specific Gravity	**13**	Pus Cell clumps
2	Albumin	**14**	Age
3	Sugar	**15**	Blood
4	Red Blood Cells	**16**	Blood Glucose Random
5	Pus Cell	**17**	Blood Urea
6	Bacteria	**18**	Serum Creatinine
7	Hypertension	**19**	Sodium
8	Diabetes Mellitus	**20**	Potassium
9	Coronary Artery Disease	**21**	Hemoglobin
10	Appetite	**22**	Packed Cell Volume
11	Pedal Edema	**23**	White Blood Cell Count
12	Anemia	**24**	Red Blood Cell Count

Figure 2.2.1: Features of others research work

Hence, we can say they have worked with very few classification techniques whereas we have worked on several classification techniques having extra features or symptoms that a patient may have. Thus makes our research much stronger and effective than others.

Procedures of their work flow are as follow:

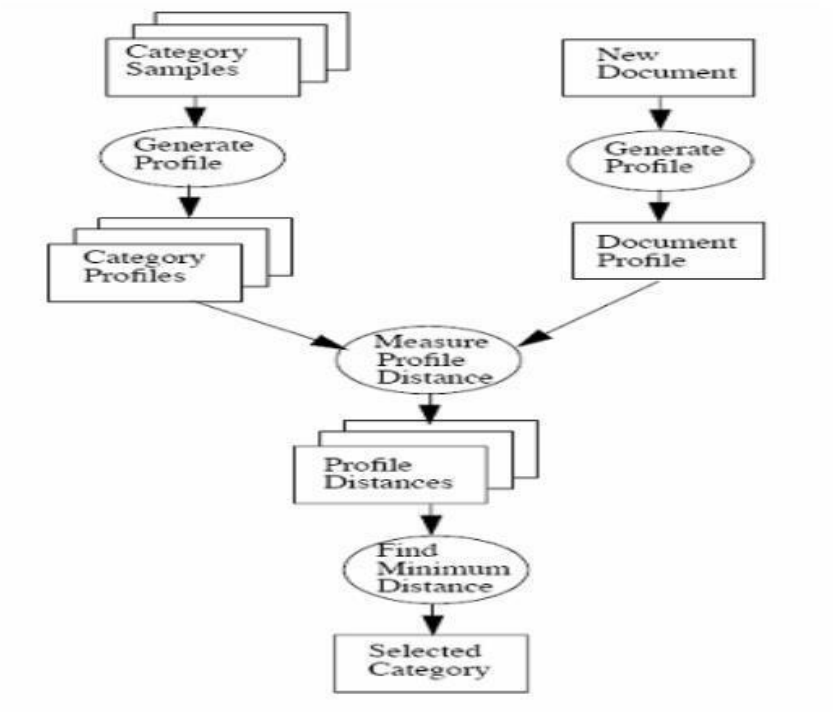


Figure 2.2.2: Classification procedure of others research

Observation

In their experiment, they found that ANN is giving better performance than any other classification techniques. They thought the reason behind giving better of performance is it could hold more information for modeling the data.

2.3 Research Summary

The above discussion done on various types of research works from different research teams, it is being appeared to us that recently, research work on Chronic Kidney Disease is increasing day by day. Some good outcomes already prove this statement well. Though, enough resources are not present, but hope is that this field is becoming more resourceful each after passing a single day.

2.3 Scope of the Problem

Kidney disease is a major health problem in Bangladesh. Day by day it is increasing in an alarming rate. There is no suitable model to justify this yet. So there is a huge scope for this problem to identify by analyzing the symptoms of kidney disease whether a patient have kidney disease or not.

2.5 Challenges

The main challenges of this work are dealing with the datasets. To clean the dataset, we need some efficient approaches to perform it but there are not enough recognized approaches to do it. Another challenge of this work is not having enough resources available regarding this topic.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter mainly deals with the theoretical knowledge of the research work. It will give the clear understanding of the concept of work. To make it more clear, very first, Research Subject and Instrumentation is explained shortly. Then we know that in the data mining or machine learning process data are the heart. For this reason, data collection process is described in this section. The chapter is being closed by giving the explanation of our project's statistical theories and besides, giving the clear concept of the implementation requirements.

3.2 Research Subject and Instrumentation

We mean by research subject is that research area that is being studied and researched for clear understandings. Not only for clear understanding, but also research subject is responsible for giving the right knowledge of various research parameters. On the other hand, Instrumentation refers to the required instruments or tools that are used by the researchers.

3.3 Data Collection Procedure

To research on specific field, the fast and foremost thing is the Data. Data is, actually, considered as the heart of the machine learning process. And for our research, there has no alternative of data. So, it has become our most challenging task for our research. We collected the global data as we could not get the real time data. We collected 400 data.

3.4 Statistical Analysis

When we deal with the raw data, the success mostly depends on the pre-processed data. The more efficiently data will be pre-processed; the outcome will be more accurate. In one word, it is the beginning challenge for such kind of research based work. Our raw data has some missing values. So it has to be included in the dataset and that was our first responsibility (Fig 3.4.1).

We have collected the global dataset which includes a patient general symptoms such as their age, blood pressure, sugar level, specific gravity, albumin, red blood cell, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia and class.

In this stage we have done data coding part to organize the data for our favor. Mainly we have organized the dataset in our way to work easily. This is how the preprocessing of dataset has been acquired. Factors that have over 15% missing information were expelled. We used some methods from pandas for filling holes or missing values in reindexed Series pad / ffill: propagate last valid observation forward to next valid backfill / bfill: use NEXT valid observation to fill gap. In this phase the data get evaluated by imputation methods and get ready for going to the next process, so that the result of the output get valid. This phase is the main part of the disease classification. Mainly, this phase decides in which way classify will be done. We used all the features for selection. So, no method was needed for this.

After successfully feature selection, we are ready for data partitioning and this is being accomplished by training our machine. We split our dataset into 3:1. The three portion of our data set are used for our training dataset and the rest portion is for testing. That means, 80% data from the datasets are used training and rest 20% is considered as the testing. In this stage, our machine is ready or fit for the classifier. We used ten different classifiers such as K-Neighbors Classifier, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Gaussian Naïve Bayes, Linear Discriminant Analysis, Logistic Regression and Artificial Neural Network to classify the data. Sklearn has built in classifier for these. We just imported it and fit it (Figure 3.4.1).

This is the final stage of our kidney disease prediction classification approach. In this stage, our model is being prepared for testing other symptoms as input data. According to the given input information, this model can classify these symptoms using several classifiers such as Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine and Random Forest.

Flowchart

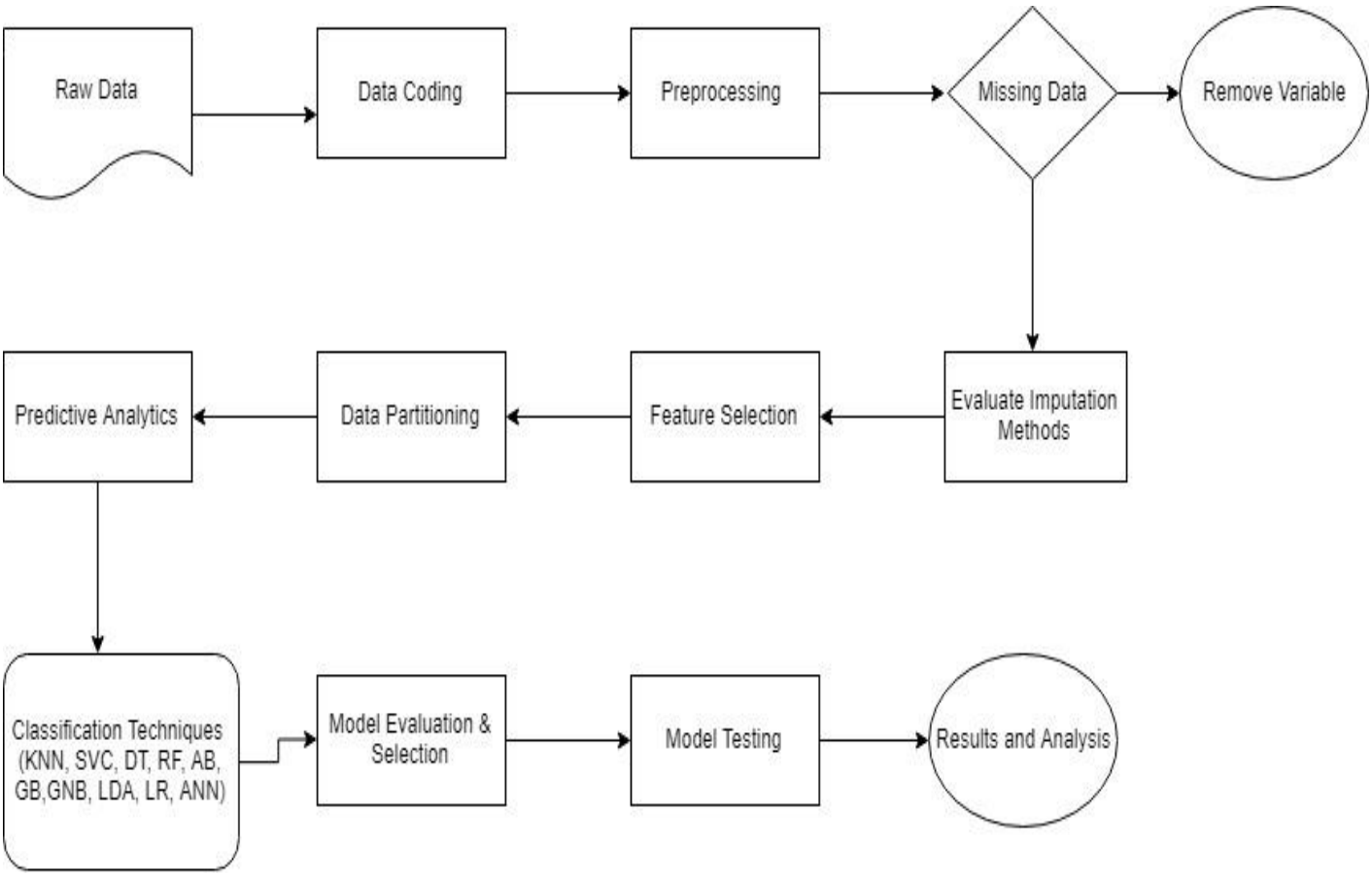


Figure 3.4.1: Flowchart for KD analysis and prediction

3.5 Implementation Requirements

After the proper analysis on all necessary statistical or theoretical concepts and methods, a list of requirement has been generated that must be required for such a work of Kidney Disease Prediction. The probable necessary things are:

Hardware/Software Requirements

- Operating System (Windows 7 or above)
- Hard Disk (minimum 4 GB)
- Ram(more than 1 GB)
- Web Browser(preferably chrome)

Developing Tools

- python 3.6
- jupyter notebook
- sklearn
- pandas
- numpy
- Keras

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This chapter 4 mainly focuses on the descriptive analysis of the data used in the research as well as experimental results of our project. Our raw data are global data. After getting the dataset, the file has been stored in csv file. In this file, there were some missing data which we had filled out by the panda's method to get value. So, in this was the dataset became handy and the pre-processed done.To build a model, we separate our dataset into two parts:

- Training Dataset
- Testing Dataset

We used 4:1 ratio for preparing our model. The four portion dataset will be treated as training dataset and the rest one portion will be considered as testing dataset.

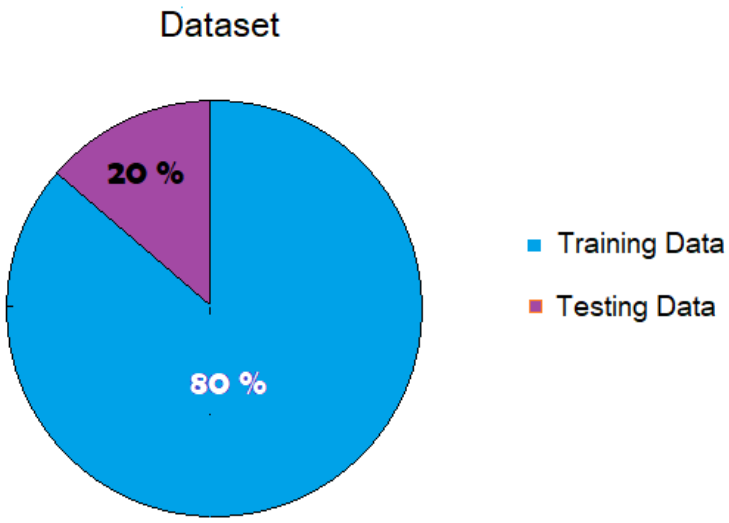


Figure 4.1.1: Dataset chart ratio

The data set for this research was obtained from online. The data set includes 400 data points and 25 attributes. There are 24 input variables and one output variable. Among the input variables, there are 11 continuous or numerical and the rest are nominal. The output variables are a class which has two categories: kd and notkd.

In order to predict KD, ten different analytics methods were used: K-Neighbors Classifier, Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Gaussian Naïve Bayes, Linear Discriminant Analysis, Logistic Regression and Artificial Neural Network. Three performance metrics are used to evaluate the analytics models: accuracy, sensitivity, and specificity. Definitions of the three metrics along with their descriptions are shown in Table 3. The confusion matrix is shown below. Positive classification is when the person has KD and negative classification of when the person does not have KD.

Metric	Description	Equation
Accuracy	Measures the ability of the model to correctly predict the class label of new or unseen data.	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	Measures the proportion of positives (or Yes's) that are correctly identified as such.	$\frac{TP}{TP + FN}$
Specificity	Measures the proportion of negatives (or No's) that are correctly identified as such.	$\frac{TN}{TN + FP}$

Abbreviation	Name	Description
TP	True Positives	Number of correct classifications predicted as positive (or Yes)
TN	True Negatives	Number of correct classifications predicted as negative (or No)
FP	False Positive	Number of examples that are incorrectly predicted as positive when it is actually negative
FN	False Negative	Number of examples that are incorrectly predicted as negative when it is actually positive

Figure 4.1.2 Accuracy, Sensitivity and Specificity

Outcome of the Diagnostic Test		Predicted	
		Positive (1)	Negative (0)
Observed	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4.1.3 Confusion Matrix

The data was partitioned into 80% for training and 20% for testing. Training data set includes 320 data points (250 with KD and 150 without KD). Testing data set includes 80 data points (49 with KD and 31 without KD). The results are shown in Table 5. For both training and testing data, the accuracy of the prediction models is very high and the model that has the highest accuracy for both data sets is Gaussian Naïve Bayes and Decision Tree classifier. This is also true for the Sensitivity and Specificity measures.

4.2 Experimental Results

Here we have discussed about the accuracy, precision, F1-score, confusion matrix and log loss of different classification techniques.

For K-Neighbor Classifier (Figure 4.2.1)

1) **Confusion Matrix:** [[22 6]
[15 37]]
2) **Accuracy :** 73.75%
3) **Sensitivity:** 0.59
4) **Specificity**: 0.86

Accuracy	73.75%
Precision	0.86
F-1 Score	0.77
Log Loss	9.06
Sensitivity	0.59
Recall	0.59

Figure 4.2.1: Accuracy of K-Neighbor Classifier

For Support Vector Classifier (SVC) (Figure 4.2.2)

1) **Confusion Matrix:** [[0 28]
   [0 52]]
2) **Accuracy :** 65.00%
3) **Sensitivity:** 0
4) **Specificity:** 0.65

Accuracy	65.00%
Precision	0.65
F-1 Score	0.78
Log Loss	12.08
Sensitivity	0
Recall	0

Figure 4.2.2: Accuracy of Support Vector Classifier (SVC)

For Decision Tree Classifier (DT) (Figure 4.2.3)

1) **Confusion Matrix:** [[28 0]
   [0 52]]
2) **Accuracy :** 100.00%
3) **Sensitivity:** 1.0
4) **Specificity:** 1.0

Accuracy	100.00%
Precision	1.0
F-1 Score	1.0
Log Loss	9.99
Sensitivity	1.0
Recall	1.0

Figure 4.2.3: Accuracy of Decision Tree Classifier (DT)

For Random Forest Classifier (RF) (Figure 4.2.4)

1) **Confusion Matrix:** [[28 0]
   [1 51]]
2) **Accuracy :** 98.75%
3) **Sensitivity:** 0.96
4) **Specificity:** 1.0

Accuracy	98.75%
Precision	1.0
F-1 Score	0.99
Log Loss	0.43
Sensitivity	0.96
Recall	0.96

Figure 4.2.4: Accuracy of Random Forest Classifier (RF)

For Ada Boost Classifier (Figure 4.2.5)

1) **Confusion Matrix:** [[28 0]
   [1 51]]
2) **Accuracy :** 98.75%
3) **Sensitivity:** 0.96
4) **Specificity:** 1.0

Accuracy	98.75%
Precision	1.0
F-1 Score	0.99
Log Loss	0.43
Sensitivity	0.96
Recall	0.96

Figure 4.2.5: Accuracy of Ada Boost Classifier

For Gradient Boosting Classifier (Figure 4.2.6)

1) **Confusion Matrix:** [[28 0]
   [1 51]]
2) **Accuracy :** 98.75%
3) **Sensitivity:** 0.96
4) **Specificity:** 1.0

Accuracy	98.75%
Precision	1.0
F-1 Score	0.99
Log Loss	0.43
Sensitivity	0.96
Recall	0.96

Figure 4.2.6: Accuracy of Gradient Boosting Classifier

For Gaussian Naïve Bayes Classifier (Figure 4.2.7)

1) **Confusion Matrix:** [[28 0]
   [0 52]]
2) **Accuracy :** 100.00%
3) **Sensitivity:** 1.0
4) **Specificity:** 1.0

Accuracy	100.00%
Precision	1.0
F-1 Score	1.0
Log Loss	9.99
Sensitivity	1.0
Recall	1.0

Figure 4.2.7: Accuracy of GaussianNB Classifier

For Linear Discriminant Analysis (Figure 4.2.8)

1) **Confusion Matrix:** [[28 0]
   [2 50]]
2) **Accuracy :** 97.50%
3) **Sensitivity:** 0.93
4) **Specificity:** 1.0

Accuracy	97.50%
Precision	1.0
F-1 Score	0.98
Log Loss	0.86
Sensitivity	0.93
Recall	0.93

Figure 4.2.8: Accuracy of Linear Discriminant Analysis

For Logistic Regression Classifier (Figure 4.2.9)

1) **Confusion Matrix:** [[27 1]
   [1 51]]
2) **Accuracy :** 97.50%
3) **Sensitivity:** 0.96
4) **Specificity:** 0.98

Accuracy	97.50%
Precision	0.98
F-1 Score	0.98
Log Loss	0.86
Sensitivity	0.96
Recall	0.96

Figure 4.2.9: Accuracy of Logistic Regression Classifier

For Artificial Neural Network Classifier (ANN) (Figure 4.2.10)

1) **Confusion Matrix:** [[0 28]
   [0 52]]
2) **Accuracy :** 65%
3) **Sensitivity:** 0
4) **Specificity:** 0.65

Accuracy	65%
Precision	0.65
F-1 Score	0.78
Log Loss	5.57
Sensitivity	0
Recall	0

Figure 4.2.10: Accuracy of ANN Classifier

4.3 Descriptive Analysis

Comparing all ten techniques we finally got the perfect techniques which satisfied the highest performance and those are decision tree and Gaussian Naïve Bayes classifier having 100% accuracy rate (Fig 4.3.2) . The data used in this study includes only two classes for the output variable, kd and notkd.

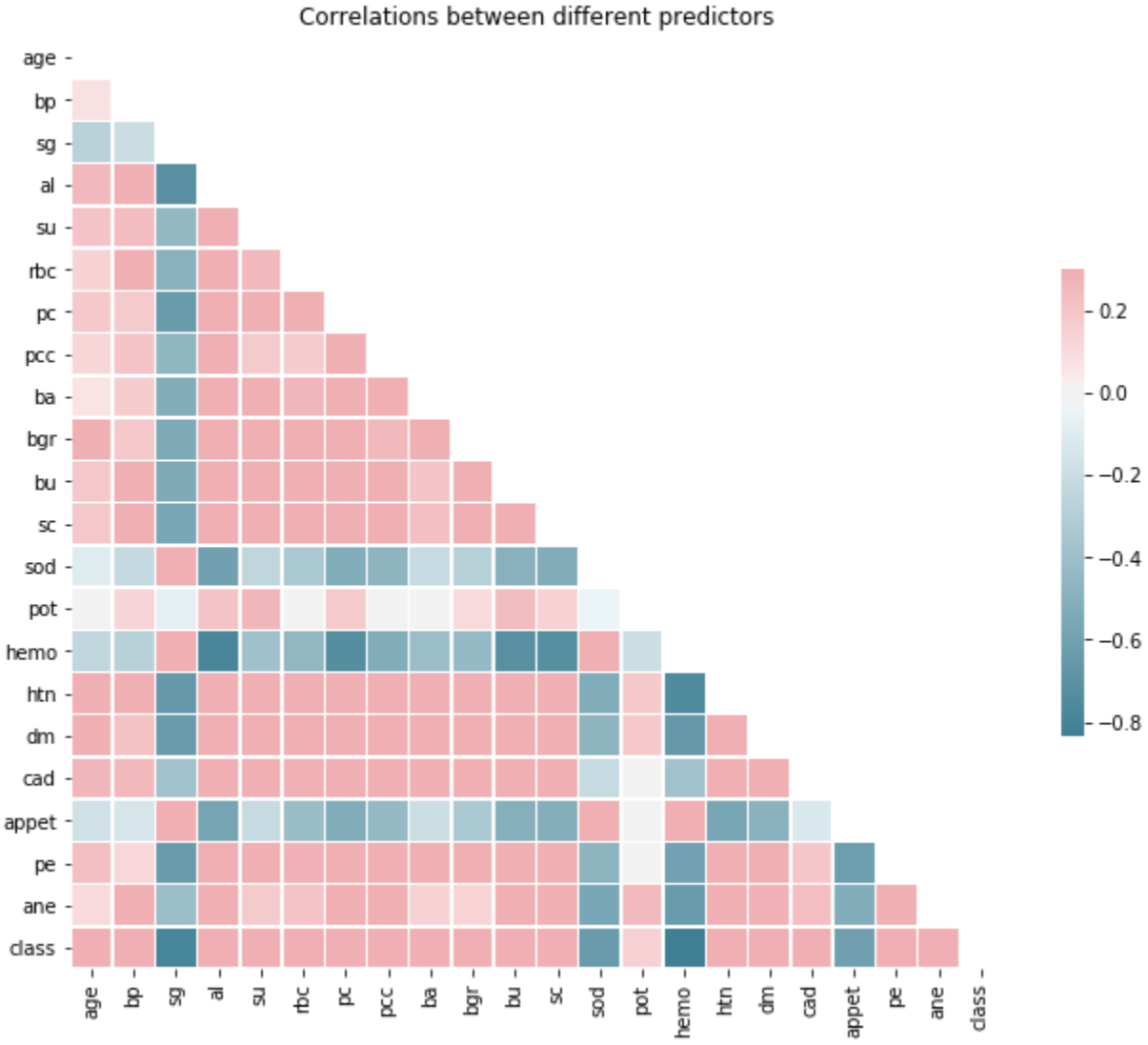


Figure 4.3.1: Correlations between different predictors

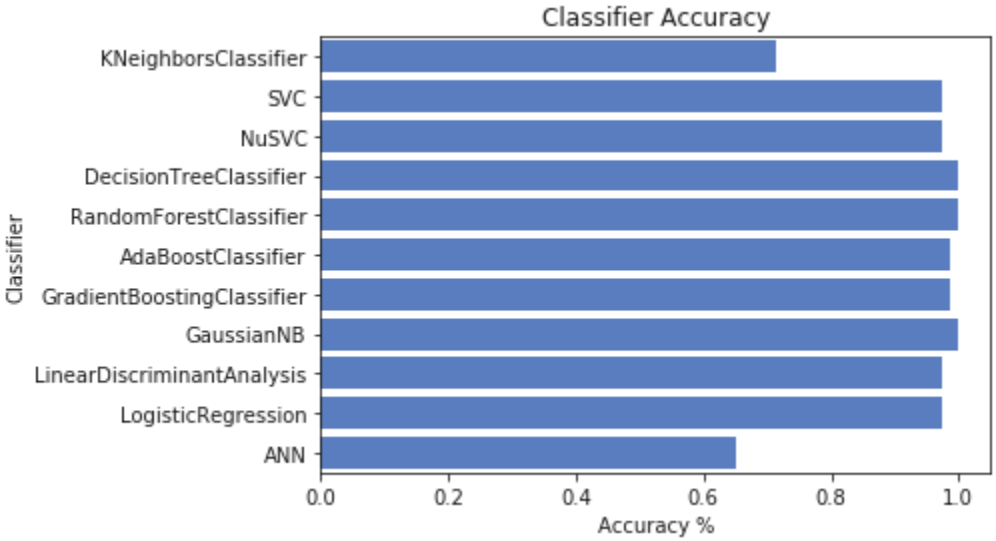


Figure 4.3.2: Classifier Accuracy

4.4 Summary

After getting this accuracy, highest result came from Decision Tree and Gaussian Naïve Bayes that are why, we are satisfied; if we try to increase accuracy level, must to prepare the dataset properly. The all categorical symptoms should be equally numbered. At that, to increase the accuracy level, data cleaning has not alternative. The more data are preprocessed, the more accurate prediction will be shown by this classifier.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AD IMPLECATION FOR FUTURE RESEARCH

5.1 Summary of the Study

Our point is to build up an upgraded and proficient machine learning (ML) application that can viably perceive and anticipate the state of interminable kidney ailment. In this work, ten most essential machine learning arrangement procedures were considered for foreseeing perpetual kidney sickness. In this procedure, the information has been partitioned into two segments. In one area train dataset got prepared and another segment got assessed by test dataset. The examination results demonstrate that Decision Tree Classifier and Gaussian Naïve Bayes accomplished most astounding execution than alternate classifiers, acquiring the F1 proportion of 1.0.

5.2 Conclusion

The accuracy level of the classifier algorithm that we used in our project is as good as we wanted. After completing this, we can say that we have learnt lots of things from this research. We can now deal with the dataset to be trained. We can now preprocess the raw data and can apply the classifier on our trained dataset. Hope, it will be very beneficial to the future researchers to do such kind of research on Kidney Disease.

5.3 Recommendations

A few notable recommendations for this are as follows:

- To create the data set more efficiently, can produce a better output of this research work.

5.4 Implication for Further Study

- Adding more categories in this project, can make this more efficient.
- Using more classifiers on this dataset, can get a better understanding on which classifier can be the best for this work.

REFERENCES

[1] Lee, S.J., and Jeon, J.H., 2015, "Relationship between Symptom Clusters and Quality of Life in Patients at Stages 2 to 4 Chronic Kidney Disease in Korea," Applied Nursing Research, 28(4), 13-19.

[2] Bala, S., and Kumar, K., 2014, "A Literature Review on Kidney Disease Prediction Using Data Mining Classification Technique," International Journal of Computer Science & Mobile Computing, 3(7), 960-967.

[3] Tangari, N., Kitsios, G.D., et al., 2013, "Risk Prediction Models for Patients with Chronic Kidney Disease" Annals of Internal Medicine, 158(8), 596-603.

[4] Hippisley-Cox, J., and Coupland, C., 2010, "Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the QKidney® Scores," Hippisley-Cox and Coupland BMC Family Practice, 11-49.

[5] Ziyad, A., 2013, "Prediction of Renal End Points in Chronic Kidney Disease," Kidney International, 83(2), 189-191.

[6] Kostoff, R.N., and Patel, U., 2015, "Literature-Related Discovery and Innovation: Chronic Kidney Disease," Technological Forecasting and Social Change, 91, 341-351.

[7] Hernandez-Pereira, E., Alvarez-Estevez, D., and Moret-Bonillo, V., 2015, "Automatic Classification of Respiratory Patterns Involving Missing Data Imputation Techniques," Biosystems Engineering, 138, 65-76.

[8] Tangri, N., Stevens, L., et al., 2011, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure," Journal of American Medical Association,305 (15), 1553-1559.

[9] Kumar, K., and Abhishek, 2012, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal of Information Technology and Computer Science, 7, 20-25.

[10] Lakshmi, K.R., Nagesh, Y., and VeeraKrishna, M., 2014, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering and Technology, 7(1), 242-254.

[11] Vijayarani, S., and Dhayanand, S., 2015, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics and Information, 4(4), 13-25.

[12] Kobayashi, T., Yoshida, T., et al., 2014, "A Metabolomics-Based Approach for Predicting Stages of Chronic Kidney Disease," Biochemical and Biophysical Research Communications, 445, 412–416.

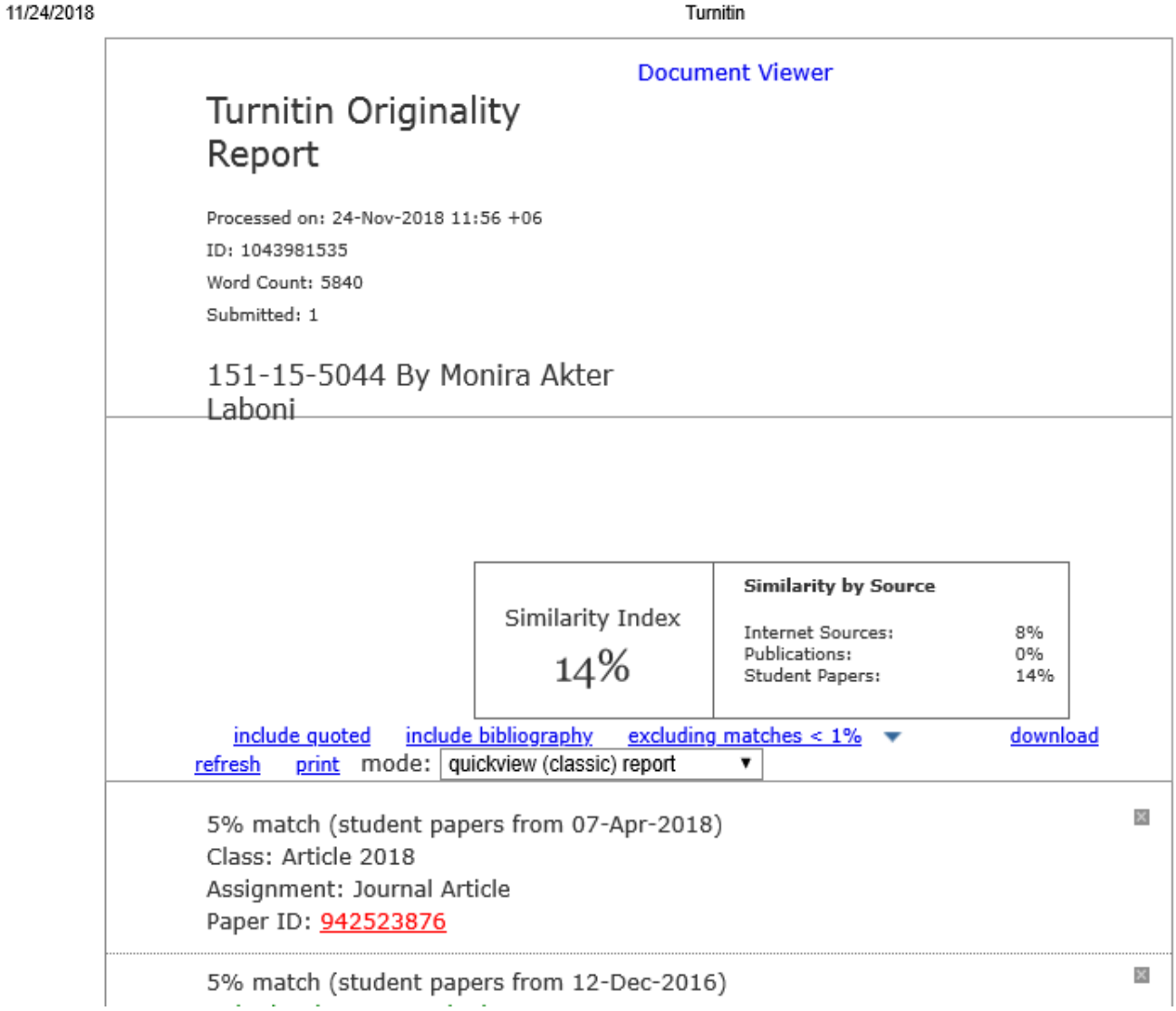
[13] Noia, T.D., Ostuni, V.C., et al., 2013, "An End Stage Kidney Disease Predictor Based on Artificial Neural Networks Ensemble," Expert Systems with Applications, 40, 4438-4445.

APPENDIX

Project Reflection

To complete the project we faced so many problem, first one was to determine the methodological approach for our project. It was not traditional work it was a research based project. We could not get that much good dataset from anywhere. Another problem was that, collection of data; it was big challenge for us. There was no available good source where we could get kidney disease dataset that is why we took the global dataset. After a long time with hard work we have successfully done this.

PLAGIARISM REPORT



11/24/2018 Turnitin

Document Viewer

Turnitin Originality Report

Processed on: 24-Nov-2018 11:56 +06
ID: 1043981535
Word Count: 5840
Submitted: 1

151-15-5044 By Monira Akter Laboni

Similarity Index	Similarity by Source	
14%	Internet Sources:	8%
	Publications:	0%
	Student Papers:	14%

include quoted include bibliography excluding matches < 1% download
refresh print mode: quickview (classic) report

5% match (student papers from 07-Apr-2018)
Class: Article 2018
Assignment: Journal Article
Paper ID: 942523876

5% match (student papers from 12-Dec-2016)