

**DETECTION OF EYE DISEASES FROM SYMPTOMS ANALYSIS USING MACHINE
LEARNING**

BY

MD AHSANUL KABIR BHUIYAN

ID: 171-15-8635

KAZI SALITH UR RAHMAN

ID: 171-15-8864

AND

SK SALMAN AHMED SABBIR

ID: 171-15-9398

The report is presented in partial fulfillment of the requirements for the degree of Bachelor of Science in computer science and engineering.

Supervised By

AHMED AL MAROUF

Lecturer

Department of CSE

Daffodil International University

Co-supervised By

SHAH MD TANVIR

SIDDIQUEE

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2021

APPROVAL

This Project/internship titled “**Detection of Eye Diseases from Symptom Analysis Using Machine Learning**”, submitted by Ahsanul Kabir Bhuiyan, Kazi Salith Ur Rahman, SK Salman Ahmed Sabbir, ID No: 171-15-8635, 171-15-8864, 171-15-9398 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 and approved as to its style and contents. The presentation has been held on 31 January 2021.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Chairman

Moushumi Zaman Bonny
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Sazzadur Ahamed
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Dr. Md Arshad Ali
Associate Professor

Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

External Examiner

DECLARATION

We hereby declare that, this thesis has been done by our team under the supervision of **AHMED AL MAROUF, Lecturer, Department of 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:



Ahmed Al Marouf

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:



Shah Md. Tanvir

Siddiquee

Assistant Professor

Department of

CSE

Daffodil International University

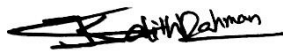
Submitted by:

Md Ahsanul kabin Bhuiyan

Md Ahsanul Kabir Bhuiyan

ID: 171-15-8635

Department of CSE



Kazi Salith Ur Rahman

ID: 171-15-9398

Department of CSE

Daffodil International University



SK Salman Ahmad Sabbir

ID: 171-15-9398

Department of CSE

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis project successfully.

We are really grateful and wish our profound indebtedness to **Ahmed Al Marouf, Lecturer, Department of CSE**, Daffodil International University, Dhaka. Deep knowledge & keen interest of our supervisor in the field of “Data Science” 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 stages have made it possible to complete this project.

We also very thankful to our domain expert, Dr. Omar Zafarullah, an ophthalmologist, Bangladesh Eye Hospital. He helps us to build our dataset in the related field.

We would like to express our heartiest gratitude to our teachers, and Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff 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 discussion while completing the course work.

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

ABSTRACT

In the area of data science, millions of data flourish every day and researchers try to use these data to make something that can help humanity and mankind. There are many machine learning approaches that are used for predicting or detecting things in many sectors. However, we work on the eye diseases prediction section in the human health sector. These days, almost all ages' people use mobile phones. Not only just use but some many of us are addicted to it. So, many of us are struggling with our eyes for it. In the whole world, many kinds of eye diseases are highly increasing day by day for many reasons especially for the reason we mention above. So, in the discussion with our domain expert, an ophthalmologist, we successfully made a dataset with several symptoms of eye diseases and the outcome of these diseases. We apply some machine learning approaches to this dataset to predict the disease with respect to the specific symptoms. We use some Meta classifiers, some tree-based algorithms, and other probabilistic algorithms for this work. We use cross-validation techniques and percentage splits techniques for better output and accuracy.

TABLE OF CONTENTS

CONTENT	PAGE
Approval	ii
Board of examiner	ii
Declarations	iii
Acknowledgement	iv
Abstract	v
 CHAPTER	
 CHAPTER – 1: INTRODUCTION	 1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the study	2
1.4 Research question	3
1.5 Expected output	3
1.6 Project management and finance	4
1.7 Report layout	4
 CHAPTER – 2: BACKGROUND	 5-7
2.1 Terminalogies	5
2.2 Scope of the problem	6
2.3 Challenges	6

CHAPTER – 3: RESEARCH METHODOLOGY	8-22
3.1 Research subject and instrument	8
3.2 Utilized dataset	8
3.3 Statistical analysis	10
3.4 Proposed methodology	20
3.5 Implementation requirements	22
 CHAPTER – 4: EXPERIMENTAL RESULTS AND DISCUSSION	 23-25
4.1 Experimental setup	23
4.2 Experimental setup and analysis	24
4.3 Discussion	25
 CHAPTER – 5: IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY	 26-28
5.1 Impact on society	26
5.2 Impact on environment	26
5.3 Ethical aspects	27
5.4 Sustainability plan	28
 CHAPTER – 6: CONCLUSION	 29-30
6.1 Summary of the study	29
6.2 Conclusion	29
6.3 Implication of future study	29
 REFERENCES	 31

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1.1: A clear view of data mining process	2
Figure 3.1: Comparison Bar chart of 3-fold Cross Validation (Algorithms)	10
Figure 3.2: Classifiers Accuracy Comparison 3-F CV:	11
Figure 3.3: Comparison Bar chart of 3-fold Cross Validation (tree based)	12
Figure 3.4: Cross Validation Accuracy (Overall)	13
Figure 3.5: Splits Accuracy (Overall)	14
Figure 3.6: Accuracy Based on Different types of algorithms K-fold CV	15
Figure 3.7: Accuracy Based on Different types of algorithms Various Splits	16
Figure 3.8: K-fold cross validation accuracy comparison	17
Figure 3.9: Various splits accuracy comparison	18
Figure 3.10: Cross validation VS splits	19

LIST OF TABLES

TABLES	PAGE NO
Table 1.1: Project Management and Timeline	4
Table 3.1: Attributes and descriptions in dataset	9
Table 4.1: Precision, Recall, F-score and Accuracy for the Algorithms used k-fold cross validation based	24
Table 4.2: Precision, Recall, F-score and Accuracy for the Algorithms used Various splits	25

CHAPTER - 1

INTRODUCTION

1.1 INTRODUCTION

The human eye is called a sensory receptor that reacts to light and also allows vision. Cone and rod cells within the retina make allowance for conscious light perception and eyesight including color differentiation and therefore the perception of depth. The eye is a component of the sensory systema nervosum. The human eyes differentiate between approximately 10 million colors and may possibly capable to detecting one photon.

Corresponding to the eyes of other mammals, the human eye's non-image-forming photosensitive ganglion cells within the retina take in light signals which sense the adjustment of the degree of the pupil, regulation, and suppressed power of the hormone melatonin, and entrainment of the body clock.

ANATOMY OF THE EYE:

This is somehow complex. The key structures of the eye are the given below:

- **Cornea:** The very clear tissue in the front of the eye
- **Lens:** Translucent disk inside the eye that focal points light rays onto the retina
- **Iris:** It is the colored part of the eye enclosing the pupil
- **Pupil:** In the iris there is a dark hole that controls the amount of light passing into the eye.
- **Optic nerve:** it connects the eye to the brain and transfers the electrical impulses formed by the retina to the visual cortex of the brain.
- **Retina:** This is a layer that lines the back of the eye, senses light, and creates electrical impulses that travel through the optic nerve to the brain
- **Macula:** A small centric level in the retina that allows us to see fine details rightly.

Eye problems might call for any and all of these parts.

There are many kinds of eye diseases. We work on 5 most common eye diseases in Bangladesh in our research. The diseases are:

- Glaucoma ACG
- Congenital Glaucoma
- Cataracts
- Bulgy Vision
- Ocular hypertension

1.2 MOTIVATION

There are many people are slacked about our health condition sometimes and do not go to a doctor for the checkup. So, our initial plan is making a system where a person can check his/her initial condition about the eyes that he/she struggled with. That's the main reason that a person has a bit knowledge about the problems of the eyes from the symptoms.

1.3 RATIONAL OF THE STUDY

In the field of machine learning, we use many kinds of data to learn a machine to deliver us some information we need. In this research, we work on a dataset that is based on eye diseases and their symptoms. We work on the 5 eye diseases we mentioned in 1.5. Our main purpose is to identify the specific diseases from these symptoms using machine learning techniques. Machine learning techniques used data mining algorithms to set up the model and sequence to find out the accuracy rate of classification, relationship prediction, and many others not even in the medical sector but also in diagnosis-based research places.

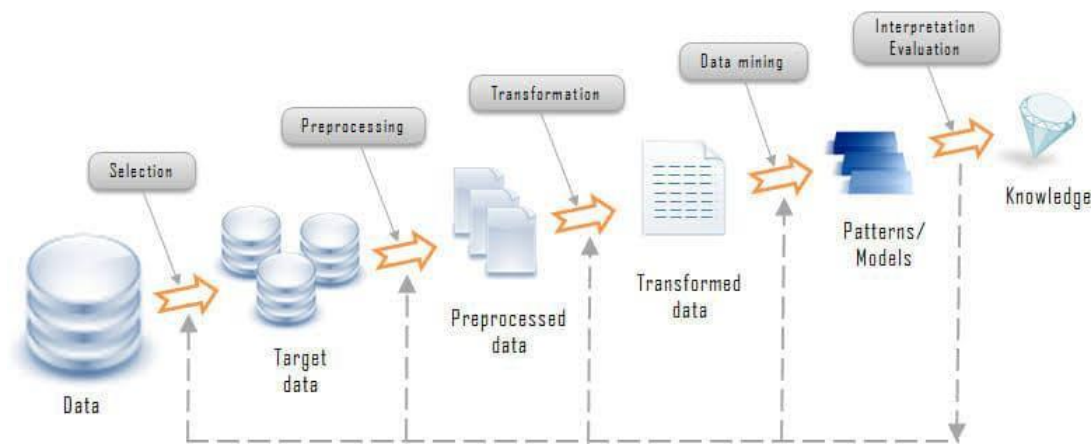


Figure 1.1: A clear view of data mining process

Machine learning is a leading process of mathematically proved algorithms and scientifically accepted statistical representations that help computer-based hardware or, software systems to carry out an appointed duty instead of using detailed instructions, relying on models and hypotheses instead. It can be announced as a part of artificial intelligence. And machine learning algorithms set up a mathematical sequence based on "training data", according to create prognostic or judgment instead of being in detail programmed to work on the task.

1.4 RESEARCH QUESTION

What is the main theme?

The main theme is detecting the eye-diseases from the symptoms.

What is the purpose of this research project?

The purpose of this project is to aware people about their eye problems. With that primary eye disease detection system, they can know about their problems and take initial steps rather than just sit in their house without any knowledge about their diseases.

Why this system is needed?

Awareness about the eye diseases and taking initial steps after knowing the problem from the system.

1.5 EXPECTED OUTPUT

In our research, we use several classification algorithms namely Random Forest Regressor, Decision Tree Classifier, KNN Classifier, Logistic Regression, Support Vector Machine, Naive Bayes Classifier, Gradient Boosting Classifier, AdaBoost Classifier, XGBoost Classifier, Multi-Class Classifier. We found many comparative views when we used these. We utilize two techniques called k-fold cross- validation and percentage splits.

There are three types of Cross-validation we use namely,

- 3-fold cross-validation
- 5-fold cross-validation
- 10-fold cross-validation

And use 3 types of percentage splits namely,

- 66% splits
- 75% splits
- 80% splits

These techniques are performed very well enough in our model and we acquire the best performance. We are comparing the algorithms also with various types of splits and cross-validation and able to acquire efficient results in many factors and circumstances.

1.6 PROJECT MANAGEMENT AND FINANCE

Table 1.1: Project Management and Timeline

Activities	Timeline
Planning	3 months
Data collection	1 month
Data processing	15 days
Implementation	20 days
Report Generation	1 month
Total	6 months 5 days

This research project is financed by the group members of this project.

1.7 REPORT LAYOUT

There are Six Chapters in our report. First chapter consists of Introduction, Motivation, Objectives, Expected Outcome, & Report Layout. Then the SECOND chapter contains Terminologies, Scope of the Problem and Challenges. The Third chapter contains Research Subject and Instrument, Data Collection Procedure, Statistical Analysis, Proposed Methodology and Implementation Requirements. Fourth Chapter contains Experimental Setup, Experimental Results and Analysis, Discussion. Fifth Chapter Consists of Society Impact, Environmental impact, Ethical Aspects, Sustainability Plan, and the last chapter includes Summary of the Study, Conclusions and Implication for Further Study.

CHAPTER – 2

BACKGROUND

2.1 TERMINALOGIES

For estimating the act of the declared Machine learning approaches, we performed some values that come from different parts. These are given below:

- **ACCURACY RATE OF CLASSIFICATION:**

The Classification accuracy Rate is measured as exactly classified samples divided by the entire number of trials multiplied by 100. Exact classified sample is the sum of True-Negative and True-Positive (TP).

Accuracy Rate is measured as, $(TP + TN / total) \times 100$.

- **PRECISION:**

Following to the Confusion Matrix, Precision is the proposition between predicted yes samples and true-positive samples.

Precision calculates as, $TP / (TP + FP)$

Here, $TP + FP = \text{Predicted Yes}$ which is followed.

- **RECALL:**

Recall is quietly known as Sensibility. Following to the Confusion Matrix, Recall is the proportion between actual yes samples and true-positive samples.

Recall calculates as, $TP / (TP + FN)$

Here, $TP + FN = \text{Actual Yes}$.

- **F-SCORE:**

F-Score is mainly called F-Measure or, F1-Score. The F-measure shall be given a more practicable measurement of a test calculation using both precision and recall. Whenever the result of F-measured becomes 1 that precises the perfection of both precision and recall.

F-Score calculates as, $(2 \times Precision \times Recall) / (Precision + Recall)$

CROSS-VALIDATION:

Cross-Validation called as an interactive heuristic activity that arbitrarily categorize the data into k-folds, each with nearly the similar number of documents, makes n-models using the alike algorithms and training parameters where every model is up skilled with n-1 folds of the data and examined on the due fold, can be attached to search the great algorithm and its top-notch optimum training parameters.

PERCENTAGE SPLIT:

It is a process of re-pattern that lay asides n% of the rows as the training dataset for constructing the model and (n-100) % of the rows retained as the test dataset to test the model. The main object classifier is trained as conversed to the trained data. On the contrary, the classification accuracy is justified on the test dataset.

2.2 SCOPE OF THE PROBLEM

Our main focus is to help people with better access to health-related information, we developed the dataset for better aggregation of eye disorders information so the users can take primary information such as what the complexity might be, what type of health experts they should consult, what preventions/precautions techniques they can take to reduce risks etc.

When we face any health issue, before consulting with a doctor-we often do a Google search or ask on social media forums to know more about the symptoms, causes, complications, and prevention of the health problem we are facing. This helps but there is a chance of getting misleading information, which may cause unnecessary mental panic and physical damage in the worst case if mistreatment is applied.

2.3 CHALLENGES

In addition, it is not a very simple task for under educated people to perform the right web search to find accurate information for the lack of technical knowledge and language barriers. However, they are used to chatting/messaging friends on social media. We will use this chat feature to supply health-related information so that people can find their desired health information by chatting as if they are messaging with a real person.

With the help of our developed eye-disorders related dataset, we are approached it with multiple machines learning techniques so that the platforms (such as Chabot, web app, etc.) consuming it can help the world with more accurate information about eye problems, causes, treatments, preventions.

In our future plan, we will develop a chatting application, which can analyze the user's message and reply with the possible information s/he is seeking, it can be found on popular IMs to be available easily without installation. Currently, it will only cover health information related to eye complications people generally face. We hope it can help users to make better decisions when facing any eye sickness.

CHAPTER - 3

RESEARCH METHODOLOGY

3.1 RESEARCH SUBJECT AND INSTRUMENT

We used 3 types of cross-validation techniques: 3-fold, 5-fold and 10-fold, and also 3- types of percentage split techniques: 66% Split, 75% Split and 80% Split. We work on our dataset with the help of our domain expert, an ophthalmologist, as our domain is “Eye disease prediction”.

3.2 UTILIZED DATASET

As our domain is eye disease so we contact an ophthalmologist as a domain expert to create a dataset about the eye diseases information. The dataset is a fully unique dataset as we are not found any work about this field or this kind of symptoms-based eye diseases prediction system. In our dataset 19 attributes are contributed as the major and minor symptoms of eye disease. There are 3 biomarkers also here. There are:

Have eye problem in family
40+ age
Diabetics

We also consider them as minor symptoms. The symptoms are considered as binary numbers because we only need to know either a specific symptom is triggered or not in a sample. The last column of our dataset indicates the result or outcome of this work which is mainly the disease's name. There are 564 active samples in this dataset.

Table 3.1: Attributes and descriptions in dataset

Attributes	Description
Cloudy, blurry or foggy vision	The values are either 0 or 1. This means this symptom have or have not
Pressure in Eye?	The values are either 0 or 1. This means this symptom have or have not
Injury to the Eye	The values are either 0 or 1. This means this symptom have or have not
Excessive dryness	The values are either 0 or 1. This means this symptom have or have not
Red eye	The values are either 0 or 1. This means this symptom have or have not
Cornea increase in size	The values are either 0 or 1. This means this symptom have or have not
Color Identifying Problem	The values are either 0 or 1. This means this symptom have or have not
Double Vision	The values are either 0 or 1. This means this symptom have or have not
Have eye problem in family	Bio mark
40+ age	Bio mark
Diabetics	Bio mark
Myopia	The values are either 0 or 1. This means this symptom have or have not
Trouble with glasses	The values are either 0 or 1. This means this symptom have or have not
Hard to see at night	The values are either 0 or 1. This means this symptom have or have not
Visible Whiteness	The values are either 0 or 1. This means this symptom have or have not
Mass pain	The values are either 0 or 1. This means this symptom have or have not
Vomiting	The values are either 0 or 1. This means this symptom have or have not
Water drops from eyes continuously	The values are either 0 or 1. This means this symptom have or have not
Presents of light when eye lid close	The values are either 0 or 1. This means this symptom have or have not
Result/Outcome	The name of the diseases for prediction

3.3 STATISTICAL ANALYSIS

CHARTS OF COMPARISONS & ACKNOWLEDGEMENT

COMPARISON OF ACCURACY | 3-F CV | ALGORITHMS:

In this part we use 4 basic machine learning algorithms and with this bar chart we show here the accuracy comparison of these algorithms with 3-fold cross validation.

The chart shows below:

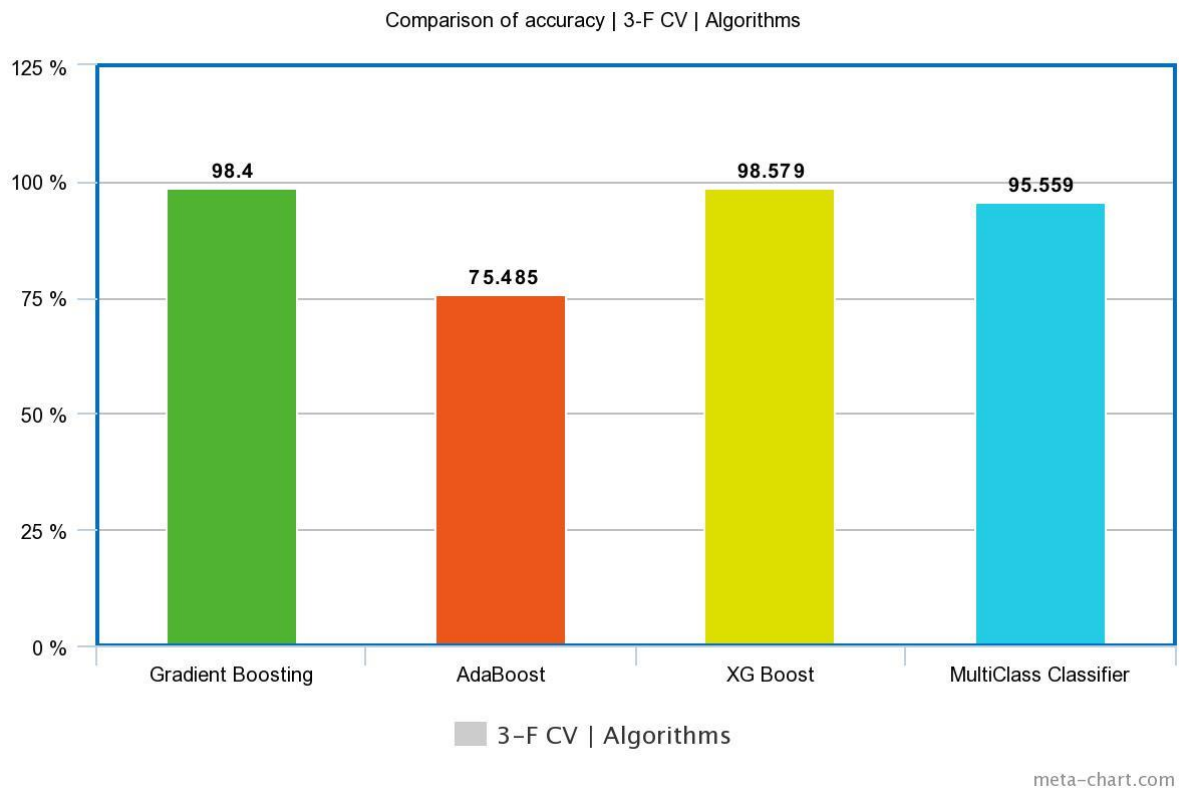


Figure 3.1: Comparison Bar chart of 3-fold Cross Validation (Algorithms)

CLASSIFIERS ACCURACY COMPARISON | 3-F CV:

Here, we use 4 meta classifiers algorithms and with this bar chart we show here the accuracy comparison of these algorithms with 3-fold cross validation.

The chart shows below:

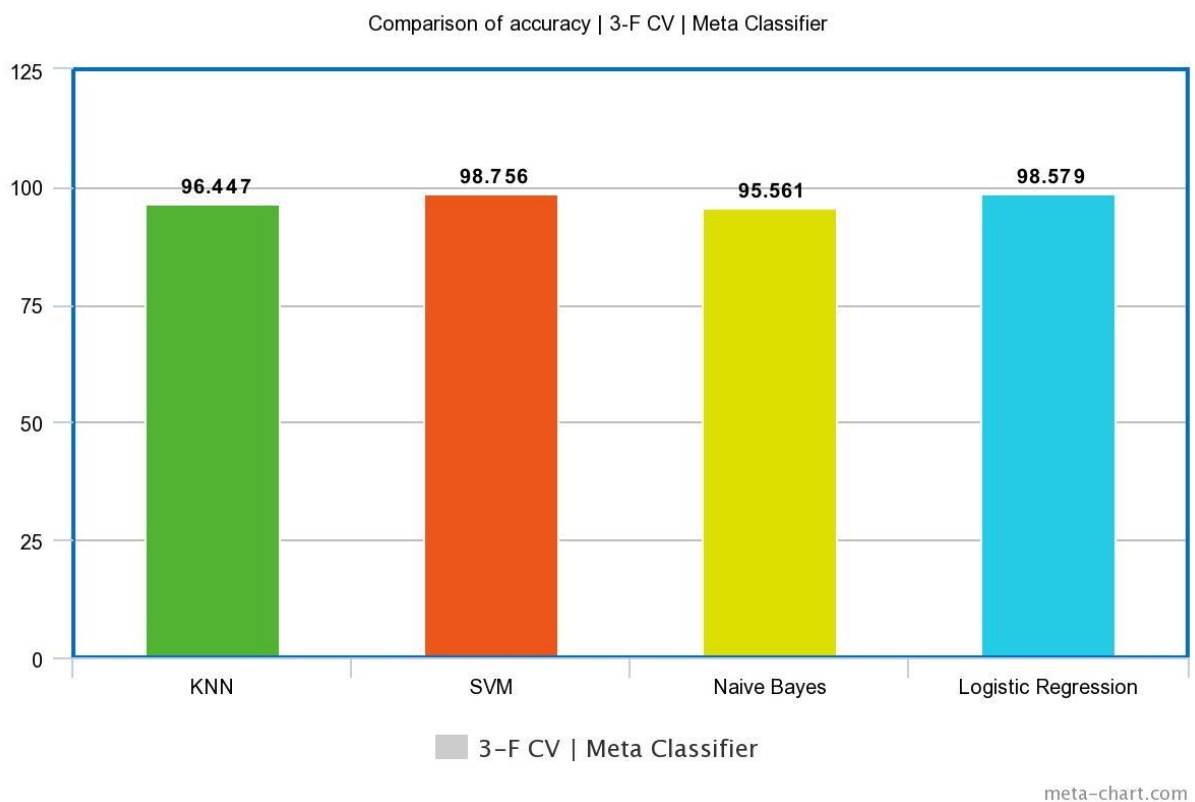


Figure3.2: Comparison Bar chart of 3-fold Cross Validation (meta classifier)

TREE BASED COMPARISON | 3-F CV:

In this part we use 2 popular machine learning tree-based algorithms and with this bar chart we show here the accuracy comparison of these algorithms with 3-fold cross validation.

The chart shows below:

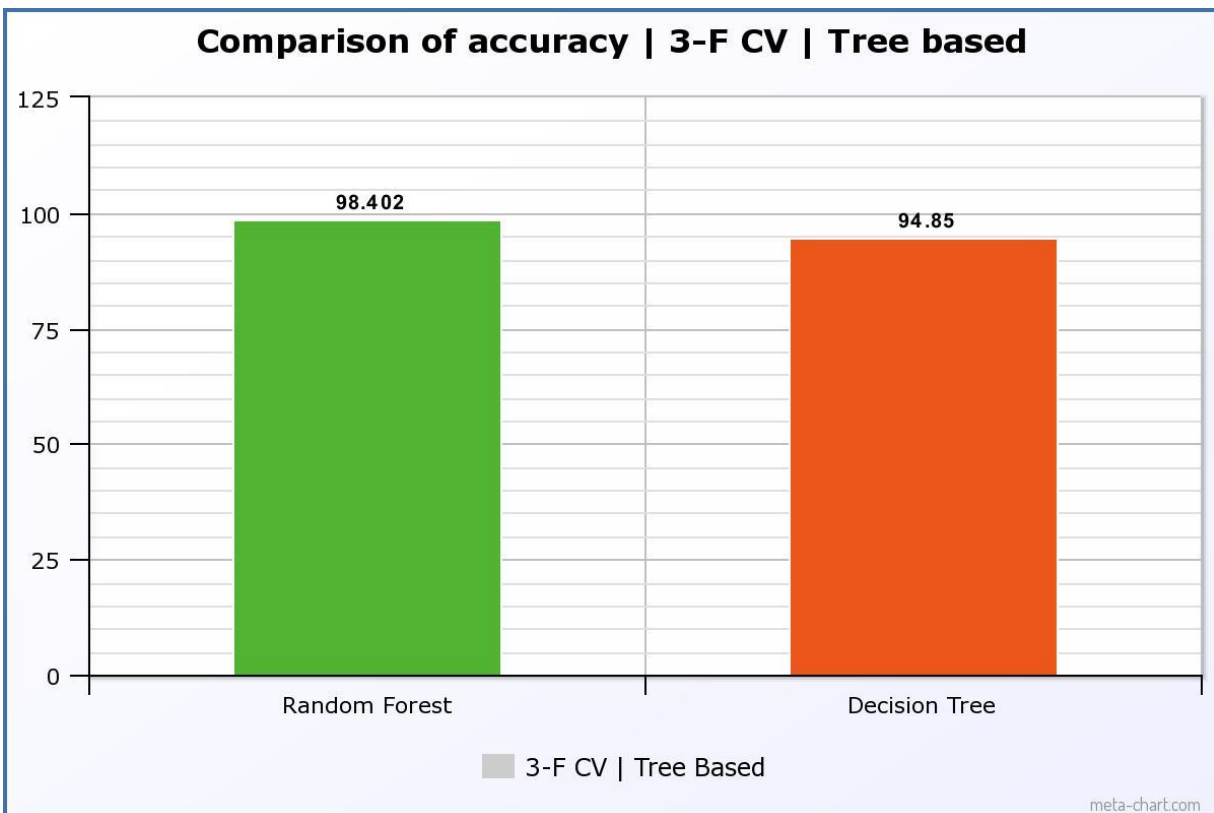


Figure 3.3: Comparison Bar chart of 3-fold Cross Validation (tree based)

OVERALL CROSS VALIDATION ACCURACY:

In this part we use all machine learning algorithms we used in our dataset and with this bar chart we show here the accuracy comparison of these algorithms with 3 basic cross validations.

The chart shows below:

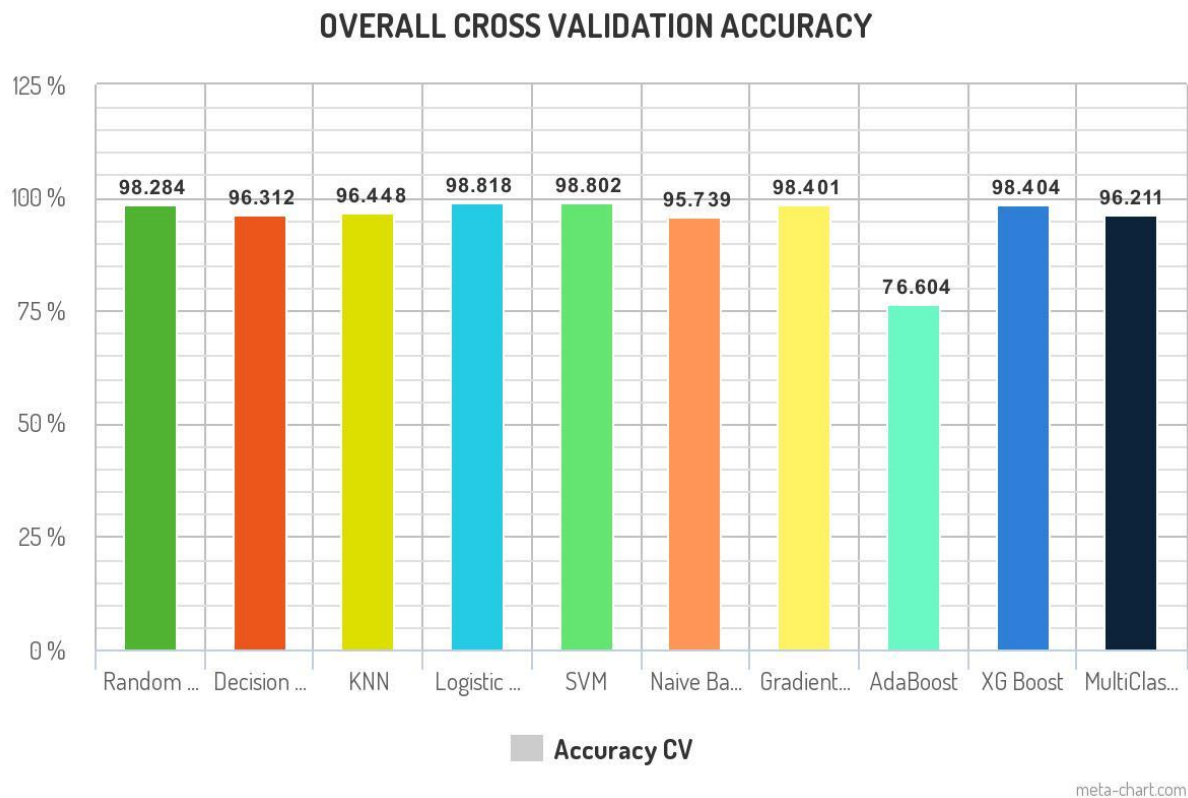


Figure 3.4: Cross Validation Accuracy (Overall)

OVERALL SPLITS ACCURACY:

In this part we use all machine learning algorithms we used in our dataset and with this bar chart we show here the accuracy comparison of these algorithms with 3 basic splits.

The chart shows below:

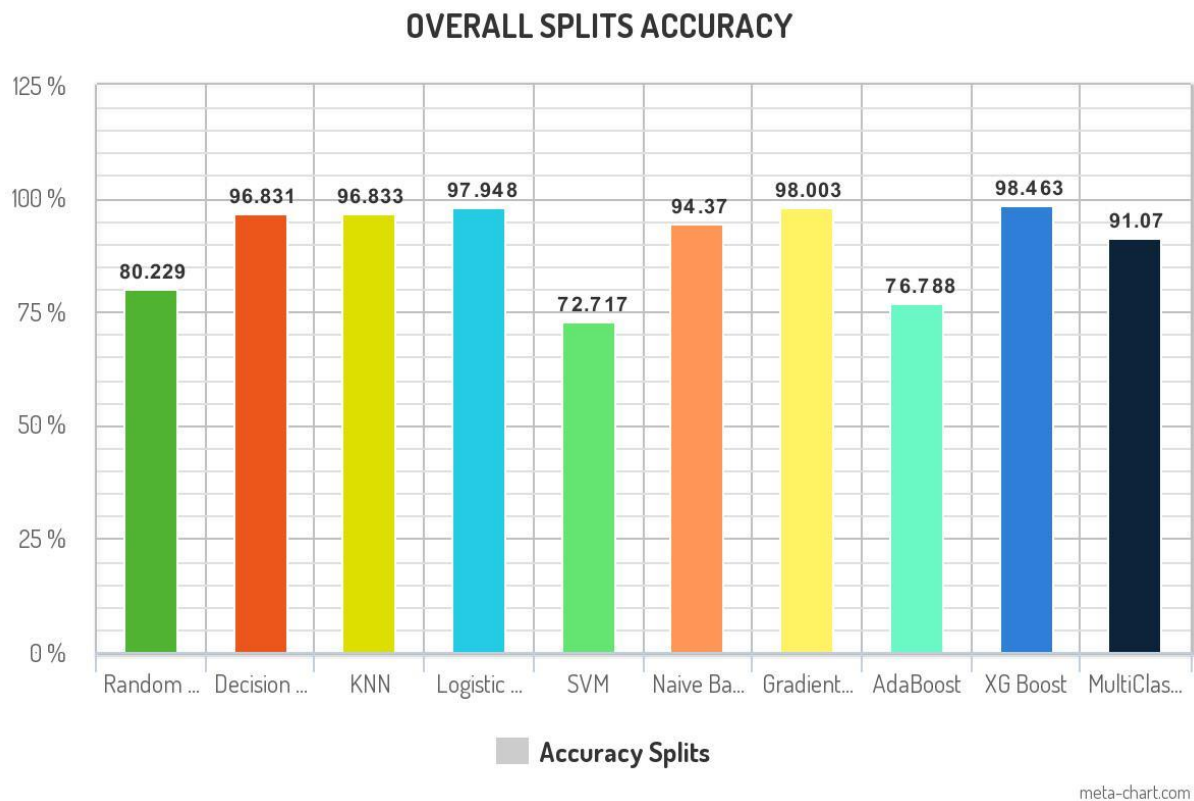


Figure 3.5: Splits Accuracy (Overall)

ACCURACY BASED ON DIFFERENT TYPES OF ALGORITHMS | K-FOLD CV:

Here, we see the difference of accuracy based on different category of algorithm in machine learning. The outcome is from k-fold Cross Validation. In this chart, we can see that the 'tree based' part has higher accuracy than other two.

The chart is given below:

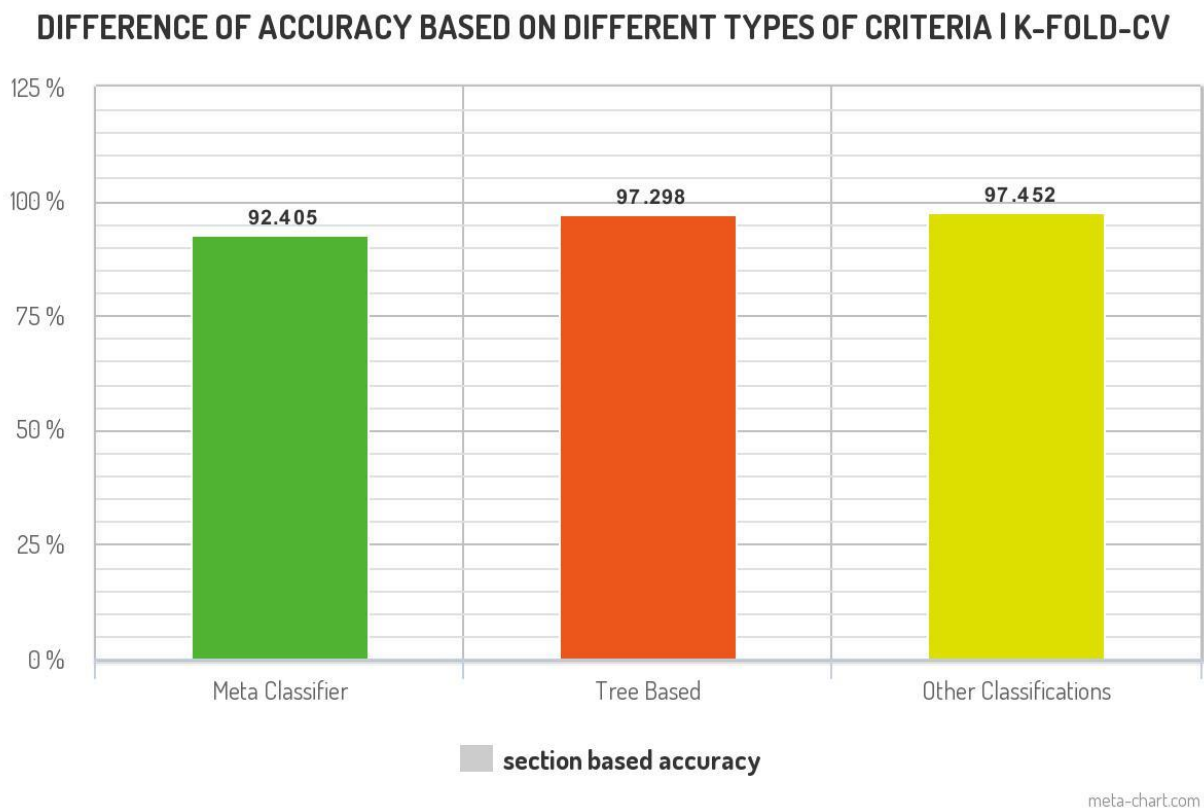


Figure 3.6: Accuracy Based on Different types of algorithms (Cross Validation)

ACCURACY BASED ON DIFFERENT TYPES OF ALGORITHM | VARIOUS SPLITS:

Here, we see the difference of accuracy based on different category of algorithm in machine learning we use in our dataset. The outcome is from various splits (66% splits, 75% split, 80% split). In this chart, we can see that the 'meta classifier' part has higher accuracy than other two.

The chart is shown below:

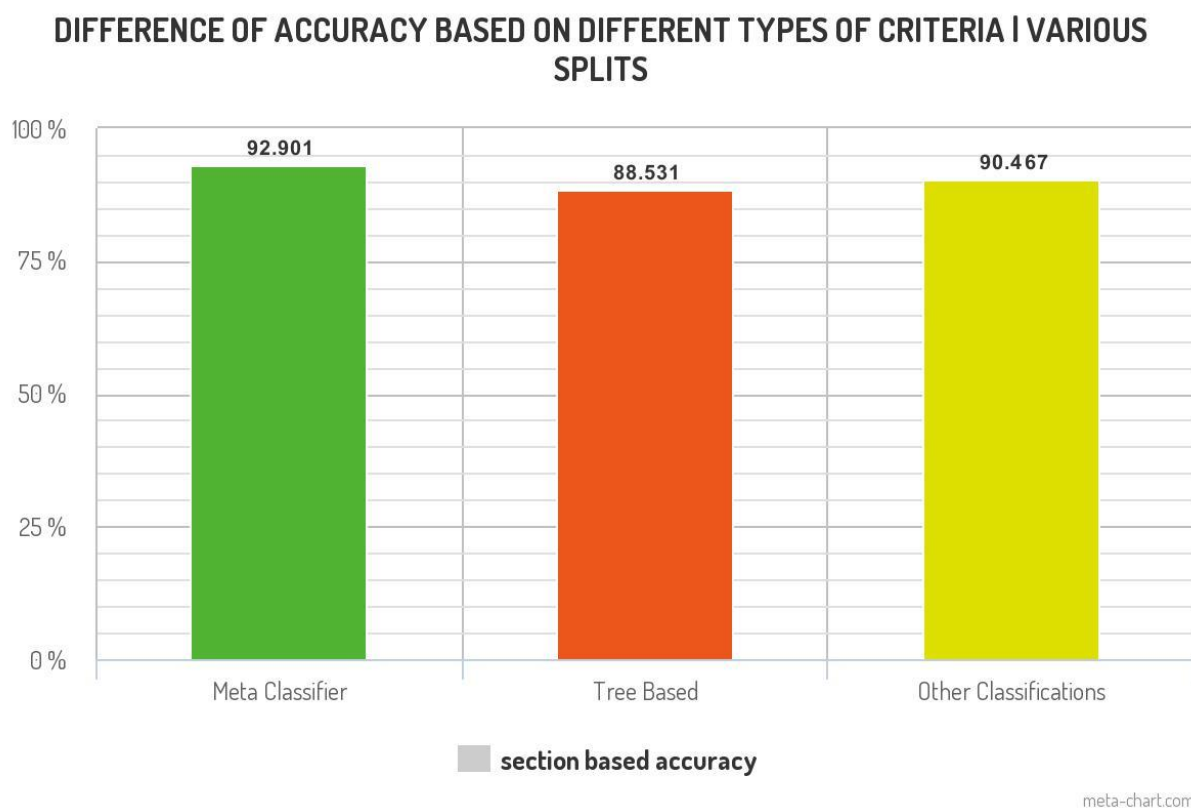


Figure 3.7: Accuracy Based on Different types of algorithm (Splits)

K-FOLD CROSS VALIDATION ACCURACY COMPARISON:

In this part, we actually compare the accuracy in different types of folds we shown in the method of cross validation. We use 10 different types of algorithms to perform these cross validations. There are 3 types of k-fold validation we used here such as 3-fold CV, 5-fold CV, 10-fold CV.

The chart is shown below:

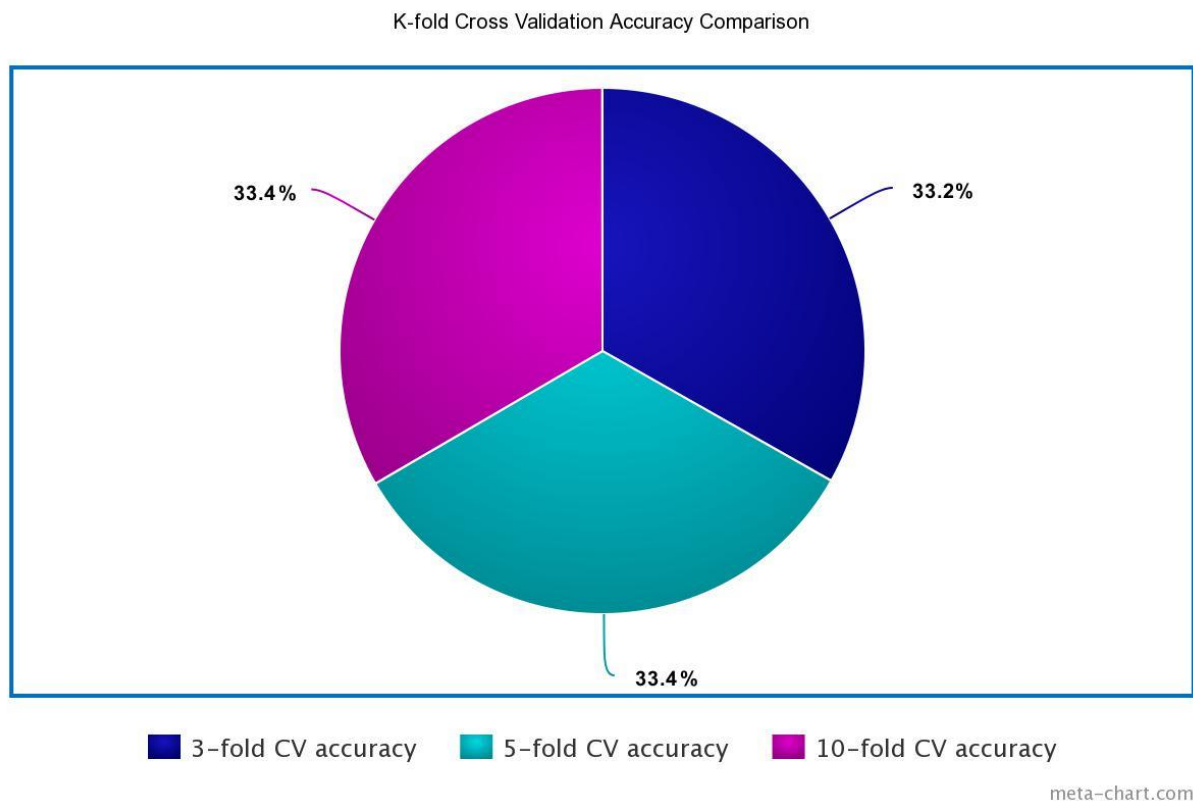


Figure 3.8: Accuracy Comparison (K-fold CV)

VARIOUS SPLITS ACCURACY COMPARISON:

In this part, we actually compare the accuracy in different types of splits we shown in this work. We use 10 different types of algorithms to perform these various. There are 3 types of splits we used here such as 66% splits, 75% splits, 80% splits.

The chart is shown below:

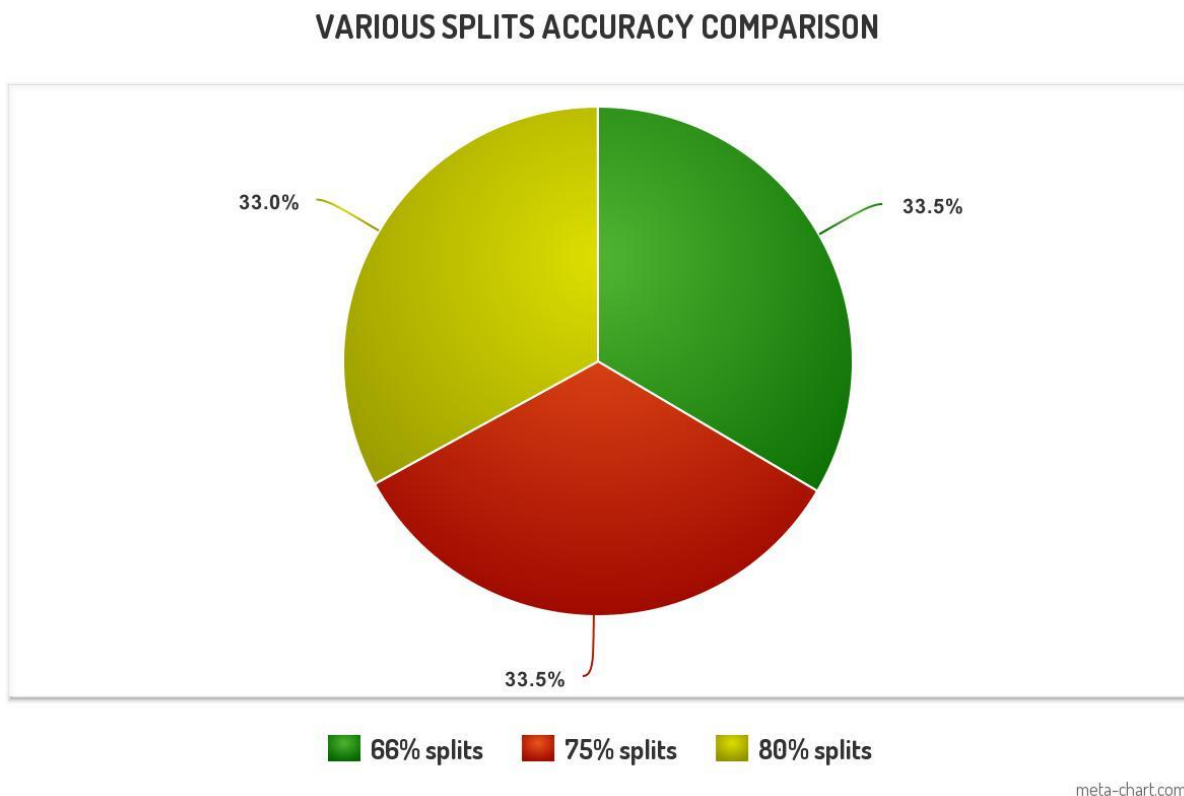
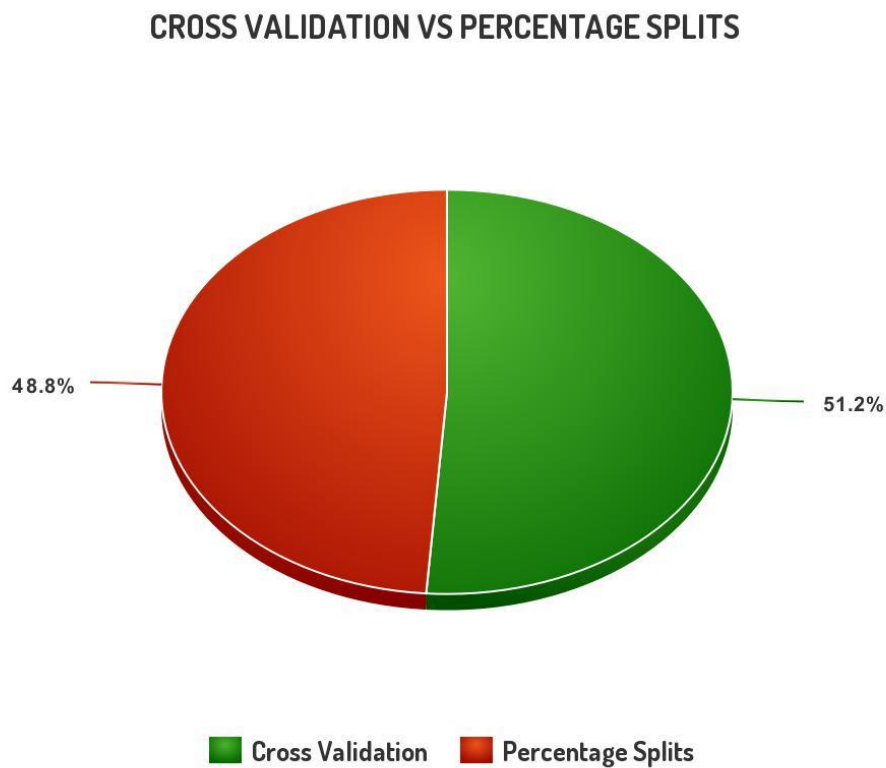


Figure 3.9: Accuracy comparison (splits)

CROSS VALIDATION VS VARIOUS SPLITS:

In this part, we compare the 2 methods we use in this work to know that which one is more efficient and gave us better result/accuracy. So, in the below chart we can see that cross- validation performs slightly well than various splits.

The chart is given below:



meta-chart.com

Figure 3.10: Cross Validation VS Splits

3.4 PROPOSED METHODOLOGY

META CLASSIFIER

□ **GRADIENT BOOSTING CLASSIFIER:**

This machine learning approach is for regression and classification, that makes a prediction representation in the form of an ensemble of low prediction models, generally decision trees. It creates the model in a stage-wise fashion alike boosting processes work, and it generalizes these by assuming optimization of an arbitrary differentiable dropping function.

□ **ADABOOST CLASSIFIER:**

It is often utilized in conjunction with many other sorts of learning algorithms to enhance performance. The opposite learning algorithms output is combined into a weighted sum that represents the ultimate output of the boosted classifier. It is adaptive within the wisdom that subsequent weak learners are tweaked in favor of these instances misclassified by previous classifiers. It is sensitive to error data and outliers. In some problems it is often less vulnerable to the over fitting problem than other learning algorithms. The individual learners are sometime weak, but as long because the performance of every one is slightly better than random guessing, the ultimate model is often proven to converge to a strong learner.

□ **XGBOOST CLASSIFIER:**

XGBoost is one of the decisions tree-based Machine Learning algorithm that utilizes a gradient boosting framework. We use this classifier in our research and it is performed well enough.

□ **MULTI-CLASS CLASSIFIER:**

In the field of machine learning, the multiclass classification is the matter of classifying examples into one of three or more classes. In this period of time many classification techniques normally admit the use of more than two classes, some are by nature binary techniques; these can, whatever, be switched into multiclass classifiers by a kinds of strategies.

TREE BASED TECHNIQUES:

□ **DECISION TREE:**

This is one of the prediction model ways used in, data mining, statistics and ML. It uses a decision tree as a predictive model to ahead from observations about an item to conclusions about the item's target value which is represented in the leaves. Tree models where the point variable can take a discrete set of values are called classification trees. In the tree structures branches represent synchronisms of features and leaves represent class labels that guide to those class labels.

□ **RANDOM FOREST REGRESSOR:**

It is called a meta estimator that conducts a number of classified decision trees on various sub-samples of the dataset and uses building to enhance the predictive accuracy and limit over-fitting

OTHER CLASSIFICATIONS

□ **KNN:**

In the field of pattern recognition, the k -nearest neighbors algorithm (KNN) is a non-parametric for classification and regression. In both facts, the input forms of the k closest training cases in the feature space. The result depends on if k -NN is used for classification or regression:

- In k -NN classification process, the result is a class community. A motive is categorized by a plurality vote of its neighbors, with the object being assigned to the class most common into its k nearest neighbors. If $k = 1$, then the item is simply engaged to the class of that single nearest neighbor.
- In k -NN regression method, the output is the property standard for the item. This value is the mean of the items of k nearest neighbors.

□ **NAIVE BAYES:**

In the field of statistics, Naive Bayes classifiers are a household of easy “probabilistic classifiers” risen on applying Bayes’ theorem with strong (naive) independence audacities between the features. They are among the easiest Bayesian network models, but paired with Kernel density estimation, they have got higher accuracy levels.

❑ **SUPPORT VECTOR MACHINE:**

In the field of machine learning, support-vector machines are supervised learning models with associated algorithms that explore data utilized for classification and regression exploration. It presents one of the most robust prediction methods, based on the field of statistical learning framework.

❑ **LOGISTIC REGRESSION:**

The term called logistic regression is a statistical representation that in its base form utilizes a logistic function to represent a binary dependent variable, though many more complex extensions remain. In the field of regression analysis, logistic regression is estimating the parameters of a logistic model.

3.5 IMPLEMENTATION REQUIREMENTS

- ❑ Windows Operating System.
- ❑ Anaconda Navigator using Python 3.7.
- ❑ Jupyter Nootebook IDE (Integrated Development Environment).
- ❑ Libraries: Numpy, Pandas for the data preprocessing.
- ❑ Libraries: ScikitLearn, Scipy for implementing different algorithms.
- ❑ WEKA

CHAPTER – 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 EXPERIMENTAL SETUP

We use 10 machine learning algorithm and two techniques called

- k-fold cross validation and
- Percentage splits in our model.

We are able to show precisions, recall, F-score of these techniques and successfully get out the accuracies of these algorithms. In this chapter we show some comparisons of these processes and methods.

4.2 EXPERIMENTAL SETUP AND ANALYSIS

Table 4.1: Precision, Recall, F-score and Accuracy for the Algorithms used k-fold cross validation based

Algorithms	3--Fold Cross Validation				5--Fold Cross Validation				10--Fold Cross Validation				Average Accuracy
	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Score</i>	<i>Accuracy</i>	
Random Forest Regressor	0.99	0.99	0.99	98.402%	1.00	1.00	1.00	98.581%	1.00	1.00	1.00	97.870%	98.284%
Decision Tree Classifier	0.96	0.95	0.95	94.850%	0.96	0.96	0.96	96.805%	0.93	0.93	0.92	96.980%	96.312%
K- Neighbors Classifier	0.96	0.96	0.95	96.447%	0.98	0.97	0.97	96.271%	1.00	1.00	1.00	96.626%	96.448%
Logistic Regression	0.99	0.98	0.95	98.579%	1.00	1.00	1.00	98.936%	1.00	1.00	1.00	98.938%	98.818%
Support Vector Machine	0.98	0.98	0.98	98.756%	1.00	1.00	1.00	98.936%	1.00	1.00	1.00	99.110%	98.802%
Naïve Bayes Classifier	0.96	0.96	0.96	95.561%	0.96	0.96	0.95	95.915%	0.95	0.93	0.39	95.742%	95.739%
Gradient Boosting Classifier	0.97	0.97	0.97	98.400%	0.98	0.97	0.97	98.401%	0.97	0.96	0.96	98.402%	98.401%
AdaBoost Classifier	0.59	0.74	0.65	75.485%	0.59	0.74	0.65	76.184%	0.62	0.77	0.68	78.142%	76.604%
XGBoost Classifier	0.98	0.98	0.98	98.579%	0.99	0.99	0.99	98.581%	0.98	0.98	0.98	98.051%	98.404%
Multi Class Classifier	0.96	0.96	0.94	95.559%	0.97	0.97	0.97	97.158%	0.96	0.96	0.96	95.915%	96.211%
Average	0.93	0.94	0.97	95.062%	0.94	0.96	0.95	95.579%	0.94	0.95	0.89	95.578%	95.401%

Table 4.2: Precision, Recall, F-score and Accuracy for the Algorithms used

Various splits

Algorithms	66% Split				75% Split				80% Split				Average Accuracy
	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy	
Random Forest Regressor	1.00	0.81	0.88	81.250%	1.00	0.82	0.89	81.560%	1.00	0.78	0.87	77.876%	80.229%
Decision Tree Classifier	0.97	0.97	0.97	96.875%	0.97	0.97	0.97	97.163%	0.97	0.96	0.96	96.460%	96.831%
K- Neighbors Classifier	0.97	0.97	0.97	96.875%	0.97	0.97	0.97	97.163%	0.96	0.96	0.96	96.460%	96.833%
Logistic Regression	0.98	0.98	0.98	97.917%	0.99	0.99	0.99	98.582%	0.97	0.97	0.97	97.345%	97.948%
Support Vector Machine	0.64	0.74	0.65	74.479%	0.63	0.74	0.64	73.759%	0.59	0.70	0.59	69.912%	72.717%
Naive Bayes Classifier	0.95	0.94	0.94	94.271%	0.96	0.95	0.95	95.035%	0.95	0.94	0.94	93.805%	94.370%
Gradient Boosting Classifier	0.98	0.98	0.98	97.917%	0.98	0.98	0.98	97.872%	0.98	0.98	0.98	98.230%	98.003%
AdaBoost Classifier	0.66	0.78	0.71	78.015%	0.66	0.78	0.71	78.014%	0.62	0.74	0.66	74.336%	76.788%
XGBoost Classifier	0.98	0.98	0.98	98.579%	0.99	0.99	0.99	98.581%	0.98	0.98	0.98	98.230%	98.463%
Multi Class Classifier	0.99	0.99	0.99	98.953%	0.98	0.98	0.98	97.872%	0.98	0.98	0.98	98.230%	98.352%
Average	0.91	0.91	0.91	91.513%	0.91	0.84	0.82	91.560%	0.80	0.90	0.89	90.088%	91.070%

4.3 DISCUSSION

After calculating all the data and performing all the algorithms, we found that every algorithms and techniques are not performed at the same level. There are some various factors depends on these. But on an average, cross validation technique performed well than percentage splits and logistic regression gave the highest accuracy in the system of cross validation.

CHAPTER – 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 IMPACT ON SOCIETY

Ensuring accurate and relevant medical information is still a big problem in this modern information age. Nowadays, for the improvement in information technology it is a lot easier to find knowledge about anything with the help of the internet. But it also introduces a problem that is too much information that is not authorized by experts, and because of less monitoring those information maybe non-related to a certain user, and relying on those information brings a new problem. This problem becomes more severe especially in cases related to medical and health information.

Our dataset and the end-user applications built based on this dataset can put a good impact on facing the problem stated above. As a result, the whole society gets a benefit in a sense that people won't have to search multiple medical blogs to find why they are facing certain symptoms, what disease they might be affected with, finally what possible treatments he can look up to.

Currently, our dataset only focuses on medical information about eye diseases, especially the most common 5 eye difficulties that people from Bangladesh face. Using our dataset and the machine models we trained up using the dataset, experts can come with new ideas to help society with new solutions. For example creating an interactive Chatbot or voice assistant that can communicate with people and note the problems they are facing, creating an easily-accessible form that takes user symptoms as input, and finally, analyze the input by our model and give a prediction of the disease they might be facing, what are the next steps they should take such as which medical or doctor they should take consultation from, etc.

Everything, as a result, can be used to solve the information anomaly problems we are facing in modern society. Which as a result impacts society by ensuring accurate and relevant information for its people.

5.2 IMPACT ON ENVIRONMENT

The environment can be of several types, as our dataset and the model is purely based on computers it doesn't seem to have any immediate benefit for the natural environment. But the applications of this data and model can have some positive impact on the natural environment.

One application of our dataset can be a voice assistant who notes down the primary complications of the patient and suggest a relevant doctor or next steps depending upon his condition.

This application replaces the need for a manual customer service representative who would manually take calls from patients one by one using a mobile phone, note his complications in paper or another database in his computer and then maybe take other steps which include more use of electronic appliances or natural resources. On the contrary, the voice assistant can do all the steps automatically, handle multiple callers at the same time, use no papers to note down, no telephone for its the side of communication, and voice assistant can live on just one server computer without needing an office and other resources that human assistant could take.

The use of fewer computers, less natural resources, and such little benefits aggregated can put a great positive impact on the environment because it's a 24/7 available and completely digital medical informatics solution.

5.3 ETHICAL ASPECTS

Ethics is an important issue especially when it is about a solution that involves IT and computers.

Methodologies used to build our dataset and to train our models are purely based on mathematical-statistical inferences developed by authors of the algorithms that we used. Predictions given by our model and the suggestions given by its applications should be taken just as possibilities and as a tool that helps the patient make better decisions for his treatment.

In another point of view, applications developed based on our dataset can help its users avoid misinformation and medical opinions of multiple blogs floating around the internet. So it can protect the users from unethical, biased motives of other information spreaders. For example, one can disregard the superstitions related to their certain condition and avoid applying non- scientific treatments that are very prevalent in the rural areas of our society.

Because technology is ever-evolving, it's never perfect and in no way it can be 100% sure of a decision that takes human analysis and multiple perspectives. Medical information is very sensitive and supplying information without caution can be life-threatening if applied. We fully understand this factor and discourage developing any application based on our dataset that markets itself as fully accurate information, or provides medicine suggestions without any reference of real medical professionals.

5.4 SUSTAINABILITY PLAN

This goes in line with the environmental aspect of the applications developed based on our dataset. The fact that automation can reduce the use of natural resources, use human involvement in monotonous tasks greatly promotes a more sustainable society.

From a financial point of view, although it seems at first glance to be harmful as it reduces the need for human assistance. But in the long run, it inspires humans to invest time in more technical studies and gain competency to develop more complex automated systems that can help society.

Our attempt to build an eye-disease prediction dataset is targeted firstly to ensure accurate health information for general users and secondly to automate the task of a medical assistant by promoting applications that can provide primary disease information. This helps both the general people of the society and the medical professional by saving time and by providing relevant information quickly without hassle.

So it has a basis of staying in service for a long time if we keep continuing the improvement of the dataset taking new perspectives from its users. In a broader sense, it promotes the sustainable development of the society by benefitting its general users and medical professionals

CHAPTER - 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 SUMMARY OF THE STUDY

This differentiation investigated the overall proficiency of the 10-machine learning algorithm namely Random Forest Regressor, Decision Tree Classifier, KNN, Logistic Regression, Support Vector Machine, Naive Bayes, Gradient Boosting, AdaBoost, XGBoost, Multi-Class Classifier for detecting eye diseases. And finally, Logistic regression and SVM perform better than others in the individual platform. On average, 75% split as a split's technique in 3 splits (66%, 75%, 80%) perform better than others. K-fold cross-validation gives better performance than various splits. But, most importantly according to the Splits technique, Meta classifiers perform well and according to the k-fold cross-validation method, Tree-based classifiers perform well. But if we consider overall algorithms and techniques, SVM & Logistic regression with 10-fold cross-validation takes place over all of them.

6.2 CONCLUSION

However, from our project, we come to a conclusion that how accurately we use the utilized dataset in the algorithms and both techniques called cross validation and percentage splits. It shows that the diseases detection is almost correctly identified by the machine learning approaches. But as a matter of concern, we should not follow it blindly because the system cannot assure it us the 100% resultant as doctors. But anyone can follow it as the 80% - 90% chances to have that particular diseases if anyone have the required symptoms and take positive steps on behalf of that diseases.

6.3 IMPLICATION OF FUTURE STUDY

When we face any health issue, before consulting with a doctor-we often do a google search or ask on social media forums to know more about the symptoms, causes, complications, and prevention of the health problem we are facing. This helps but there is a chance of getting misleading information, which may cause unnecessary mental panic and physical damage in the worst case if mistreatment is applied.

In addition, it is not a very simple task for undereducated people to perform the right web search to find accurate information for the lack of technical knowledge and language barriers. However, they are used to chatting/messaging friends on social media. We will use this chat feature to supply health-related information so that people can find their desired health information by chatting as if they are messaging with a real person.

We will develop a chatting application, which can analyze the user's message and reply with the possible information s/he is seeking, it can be found on popular IMs to be available easily without installation. Currently, it will only cover health information related to eye complications people generally face. We hope it can help users to make better decisions when facing any eye sickness.

REFERENCE

- [1] geeksforgeeks available at << <https://www.geeksforgeeks.org/random-forest-regression-in-python/> >>, last accessed on 20-10-2020 at 07:00 PM.
- [2] javatpoint, available at << <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm> >>, last accessed on 21-10-2020 at 08:00 PM.
- [3] javatpoint, available at << <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> >>, last accessed on 22-10-2020 at 10:00 AM.
- [4] javatpoint, available at << <https://www.javatpoint.com/logistic-regression-in-machine-learning> >>, last accessed on 22-10-2020 at 09:00 AM.
- [5] javatpoint, available at << <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm> >>, last accessed on 23-10-2020 at 07:00 PM.
- [6] javatpoint, available at << <https://www.javatpoint.com/machine-learning-naive-bayes-classifier> >>, last accessed on 25-10-2020 at 08:00 PM.
- [7] stackabuse, available at << <https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/> >>, last accessed on 28-10-2020 at 09:00 PM.
- [8] datacamp, available at << <https://www.datacamp.com/community/tutorials/adaboost-classifier-python> >>, last accessed on 02-11-2020 at 07:00 PM.
- [9] machinelearningmastery, available at << <https://machinelearningmastery.com/develop-first-xgboost-model-python-scikitlearn/> >>, last accessed on 06-11-2020 at 09:00 PM.
- [10] javatpoint, available at << <https://www.javatpoint.com/confusion-matrix-in-machine-learning> >>, last accessed on 05-11-2020 at 06:00 PM.
- [11] javatpoint, available at << <https://www.javatpoint.com/cross-validation-in-machine-learning> >>, last accessed on 9-11-2020 at 09:00 PM.

Eye Disease Detection

ORIGINALITY REPORT

28%

SIMILARITY INDEX

20%

INTERNET SOURCES

15%

PUBLICATIONS

22%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Daffodil International University

Student Paper

9%

2

"Implications of Meta Classifiers for Onset Diabetes Prediction", International Journal of Innovative Technology and Exploring Engineering, 2020

Publication

4%

3

[dspace.daffodilvarsity.edu.bd:8080](https://dspace.daffodilvarsity.edu.bd/8080)

Internet Source

2%

4

github.com

Internet Source

2%

5

www.ijircce.com

Internet Source

1%

6

medium.com

Internet Source

1%

7

Submitted to TechKnowledge

Student Paper

1%

8

Submitted to RMIT University

Student Paper

1%