DETECTION OF EYE DISEASES FROM SYMPTOMS ANALYSIS USING MACHINE LEARNING APPROACHES

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

MD. TAREK HABIB

Assistant Professor

Department of CSE

Daffodil International University

APPROVAL

This Project titled "**Detection of eye diseases from symptoms analysis using machine learning approaches**", submitted by ***Md. Ahsanul Kabir Bhuiyan, Kazi Salith Ur Rahman and SK Salman Ahmed Sabbir*** 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 12th May, 2020.

BOARD OF EXAMINERS

(Name) Chairman
Designation
Department of CSE [Font-12]
Faculty of Science & Information Technology
Daffodil International University

(Name) Internal Examiner
Designation
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

(Name) External Examiner
Designation
Department of -------
Jahangirnagar University

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:

MD Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

1. **Md. Ahsanul Kabir Bhuiyan**
ID: 171-15-8635
Departmaent of CSE
Daffodil International University

2. **Kazi Salith Ur Rahman**
ID: 171-15-8864
Departmaent of CSE
Daffodil International University

3. **SK Salman Ahmed Sabbir**
ID: 171-15-9398
Departmaent of CSE
Daffodil International University

Abstract

In the area of data science, millions of data flourish every day and researchers try to use these data to make something which can help the humanity and mankind. There are many machine learning approaches are used for predicting or detecting things in many sectors. However, we work on eye diseases prediction section in human health sector. In these days, almost all ages people use mobile phone. Not only just use but some many of us are addicted on it. So, many of us are struggling with their eyes for it. In the whole world many kinds of eye diseases are highly increasing day by day for many reasons specially for the reason we mention above. So, the discussion with our domain expert, an ophthalmologist, we successfully make a dataset with several symptoms of eye diseases and the outcome of these diseases. We apply many machine learning approaches on this dataset to predict the disease respect of 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 technique for the better output and accuracy.

Index

Table of contents:

Chapter - 1

Introduction

1.1 Eye:

The human eye is a sense organ that reacts to light and allows vision. Rod and cone cells in the retina allow conscious light perception and vision including color differentiation and the perception of depth. The human eye can differentiate between about 10 million colors and is possibly capable of detecting a single photon. The eye is part of the sensory nervous system.

Similar to the eyes of other mammals, the human eye's non-image-forming photosensitive ganglion cells in the retina receive light signals which affect adjustment of the size of the pupil, regulation and suppression of the hormone melatonin and entrainment of the body clock.

1.2 Anatomy of the eye:

The anatomy of the eye is complex. The main structures of the eye include the following:

- **Cornea**: clear tissue in the very front of the eye
- **Iris:** colored part of the eye surrounding the pupil
- **Pupil:** dark hole in the iris that regulates the amount of light going into the eye
- **Lens**: small clear disk inside the eye that focuses light rays onto the retina

- **Retina**: layer that lines the back of the eye, senses light, and creates electrical impulses that travel through the optic nerve to the brain
- **Macula**: small central area in the retina that allows us to see fine details clearly
- **Optic nerve:** connects the eye to the brain and carries the electrical impulses formed by the retina to the visual cortex of the brain
- **Vitreous**: clear, jelly-like substance that fills the middle of the eye

Eye problems can involve any and all of these parts. As you read through this article, you can refer to this illustration for reference.

1.3 Eye diseases:

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

1.4 Machine learning:

Chapter - 2

Algorithms used in this work

Classifiers

- **Gradient Boosting Classifier:**
  Gradient boosting is a machine learning technique
  for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function

- **AdaBoost Classifier:**
  This is a machine learning meta-algorithm formulated by Yoav
  Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

- **XGBoost Classifier:**
  XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) ... A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.

- **Multi-Class Classifier:**
  machine learning, multiclass or multinomial classification is the problem of classifying instances into one of three or more classes (classifying instances into one of two classes is called binary classification). While many classification algorithms (notably multinomial logistic regression) naturally permit the use of more than two classes, some are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies.

Tree Based Techniques:

- **Decision Tree:**
  Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

- **Random Forest Regressor:**
  A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Other Algorithms

- **KNN:**

  In pattern recognition, the *k*-nearest neighbors algorithm (KNN) is a non-parametric method proposed by Thomas Cover used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

  - In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

  - In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

- **Naïve Bayes:**

  In statistics, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with Kernel density estimation, they can achieve higher accuracy levels.

- **Support Vector Machine:**

  In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Developed at AT&T Bell Laboratories by Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997), it presents one of the most robust prediction

methods, based on the statistical learning framework or VC theory proposed by Vapnik and Chervonenkis (1974) and Vapnik (1982, 1995). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

- **Logistic Regression:**

  Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Chapter - 3

Experimental Model

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”.

Dataset:

As our domain is eye disease so we contact with an ophthalmologist as a domain expert to create a dataset about the eye diseases information. The dataset is 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 bio markers 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 diseases name. There are 564 active samples in this 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

Performance Measured Used:

For measuring the performance of the declared Meta Classifier, we used numerous values that come from different sectors.

- **Accuracy Rate of Classification:**

  Accuracy Rate of Classification is computed as exactly classified samples divided by the entire number of samples multiplied by 100. Exact classified sample is the sum of True-Positive (TP) and True-Negative.
  Accuracy Rate = $(TP+TN/\ Total) \times 100$

- **Precision:**

  According to the Confusion Matrix, Precision is the ratio between true-positive samples and predicted yes samples.
  Precision = $TP\ /\ (TP+FP)$
  Here, TP+FP = Predicted Yes

- **Recall:**

  Recall is also known as Sensitivity. According to the Confusion Matrix, Recall is the ratio true-positive samples and actual yes samples.
  Recall = $TP\ /\ (TP+FN)$
  Here, TP+FN = Actual Yes

- **F-Score:**

  F-Score is also called F1-Score or, F-Measure. The F-Score can give a more feasible measurement of a test implementation using both recall and precision. When the value of F-Score becomes 1 that indicates the perfection of both recall and precision.
  F-Score = $(2\times Precision\times Recall)\ /\ (Precision+Recall)$

Cross-Validation:

Cross-Validation is a heuristic works that arbitrarily classify the data into n-folds, each with nearly the similar number of records, makes n-models using the similar algorithms and training parameters where every model is trained with n-1 folds of the data and tested on the due fold, can be applied to search the best algorithm and its optimum training parameters.

Percentage Split:

Percentage Split is a process of re-sampling that reserves n% of the rows as the training dataset for structuring the model and (n-100) % of the rows reserved as the test dataset to test the model. The target classifier is trained as opposed to the trained data. On the other hand, the classification accuracy is measured on the test dataset.

Chapter - 4

Table with 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					
	Precision	*Recall*	*F-Score*	*Accuracy*	*Precision*	*Recall*	*F-Score*	*Accuracy*	*Precision*	*Recall*	*F-Score*	*Accuracy*	*Average Accuracy*	*Standard Deviation Of Accuracy*
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	0.96	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	0.98	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%	
Naive 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 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%	
MultiClass Classifier	0.96	0.96	0.94	95.5595%	0.97	0.97	0.97	97.1581%	0.96	0.96	0.96	95.9147%	96.211	
Average				95.062%				95.579%				95.578%	95.401	
Standard Deviation														

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

Various splits

Algorithms	66% Split				75% Split				80% Split					
	Precision	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Average Accuracy**	**Standard Deviation Of Accuracy**
Random Forest	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	0.98	0.97	0.97	97.396%	0.98	0.98	0.98	97.872%	0.98	0.97	0.97	97.345%	97.538%	
KNN	0.97	0.97	0.97	96.875%	0.97	0.97	0.97	97.1631%	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%	
SVM	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	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 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	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%	
MultiClassClassifier	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				91.565%				91.631%				90.177%	91.124%	
Standard Deviation														

Chapter - 5

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:

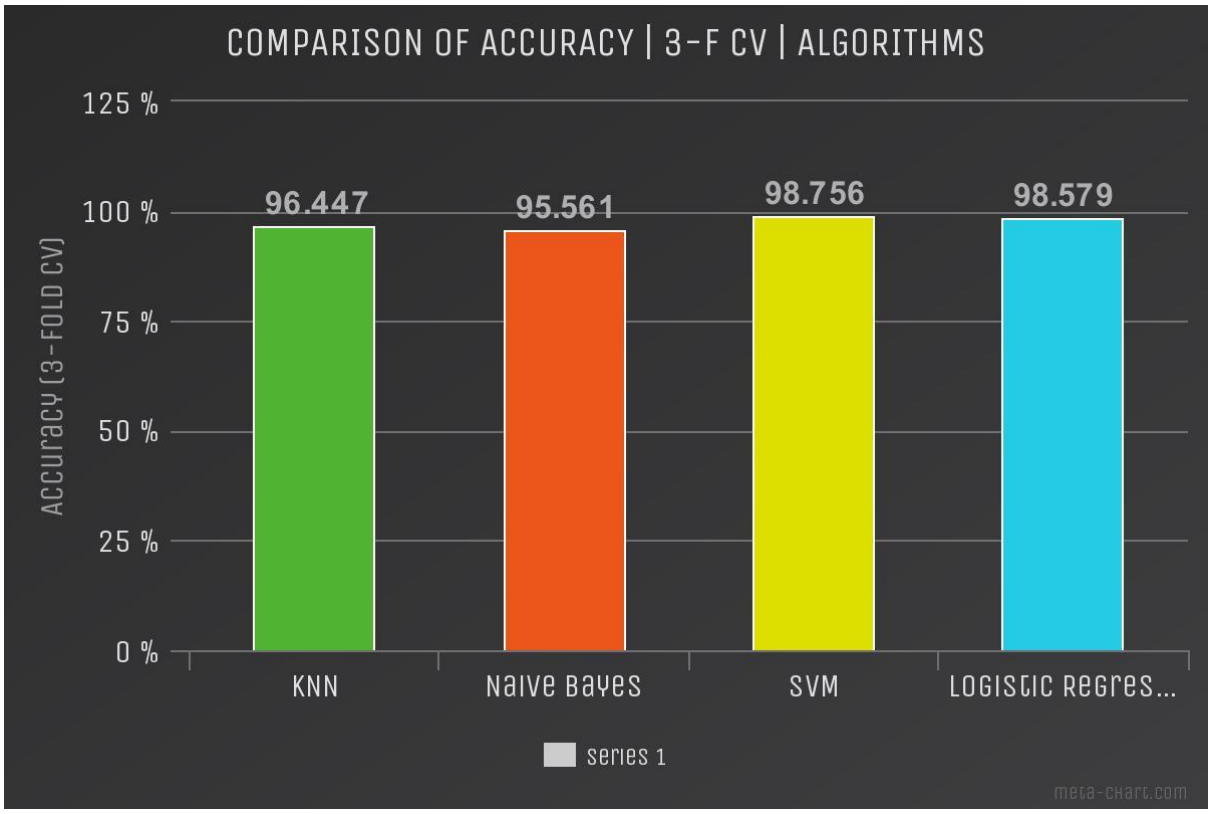

Figure: Comparison Bar chart of 3-fold Cross Validation (Algorithms)

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:

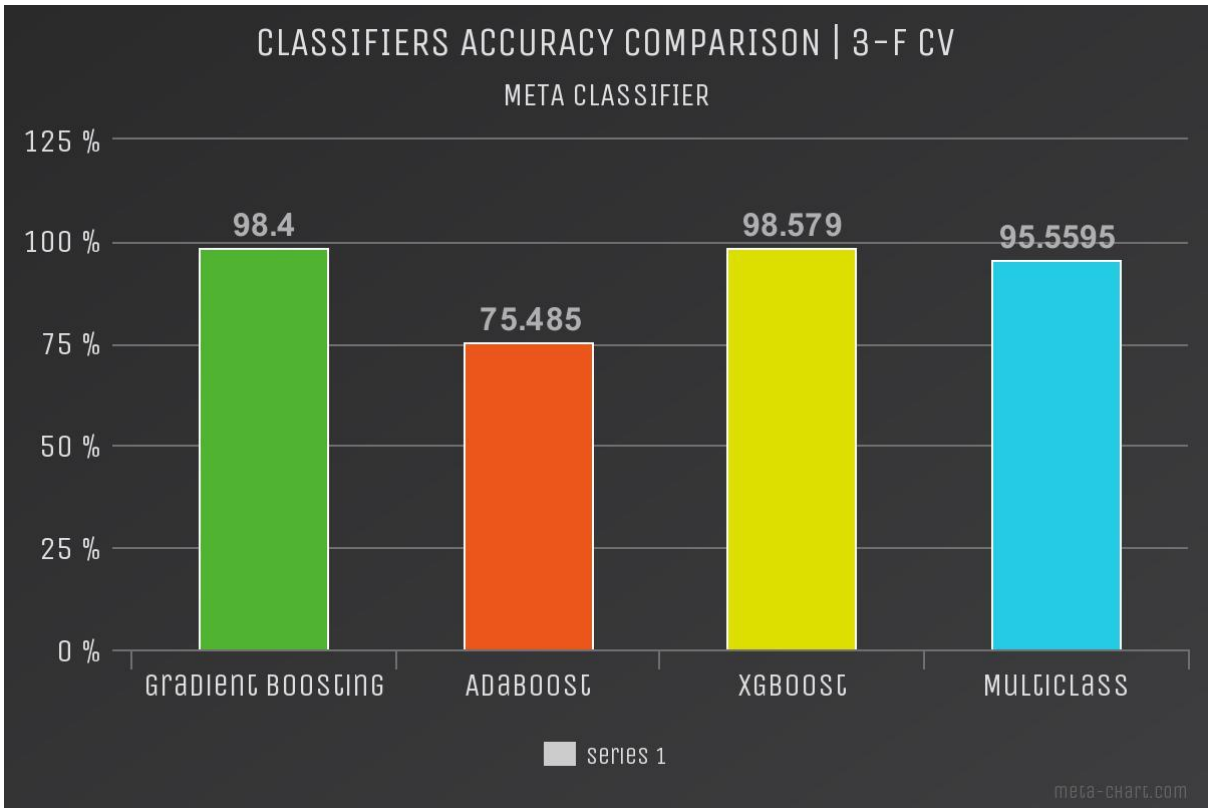

Figure: Comparison Bar chart of 3-fold Cross Validation (meta classifier)

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:

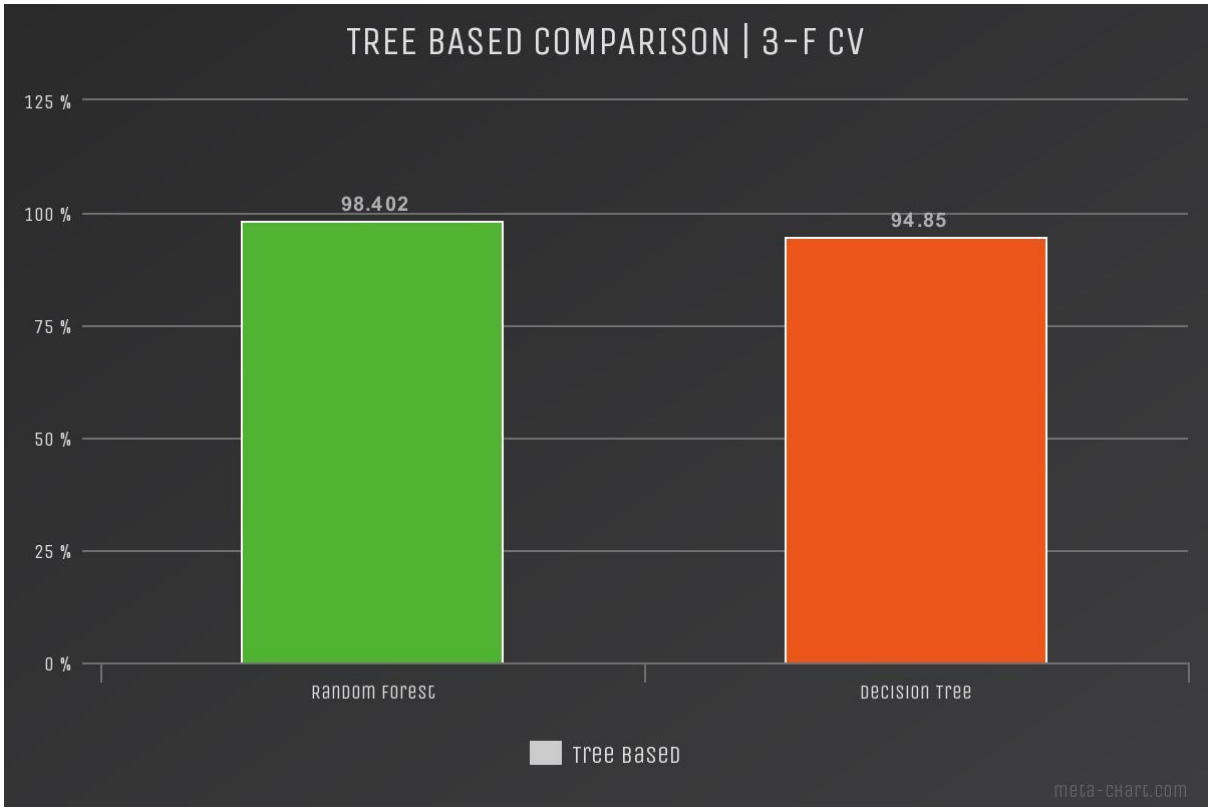

Figure: Comparison Bar chart of 3-fold Cross Validation (tree based)

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:

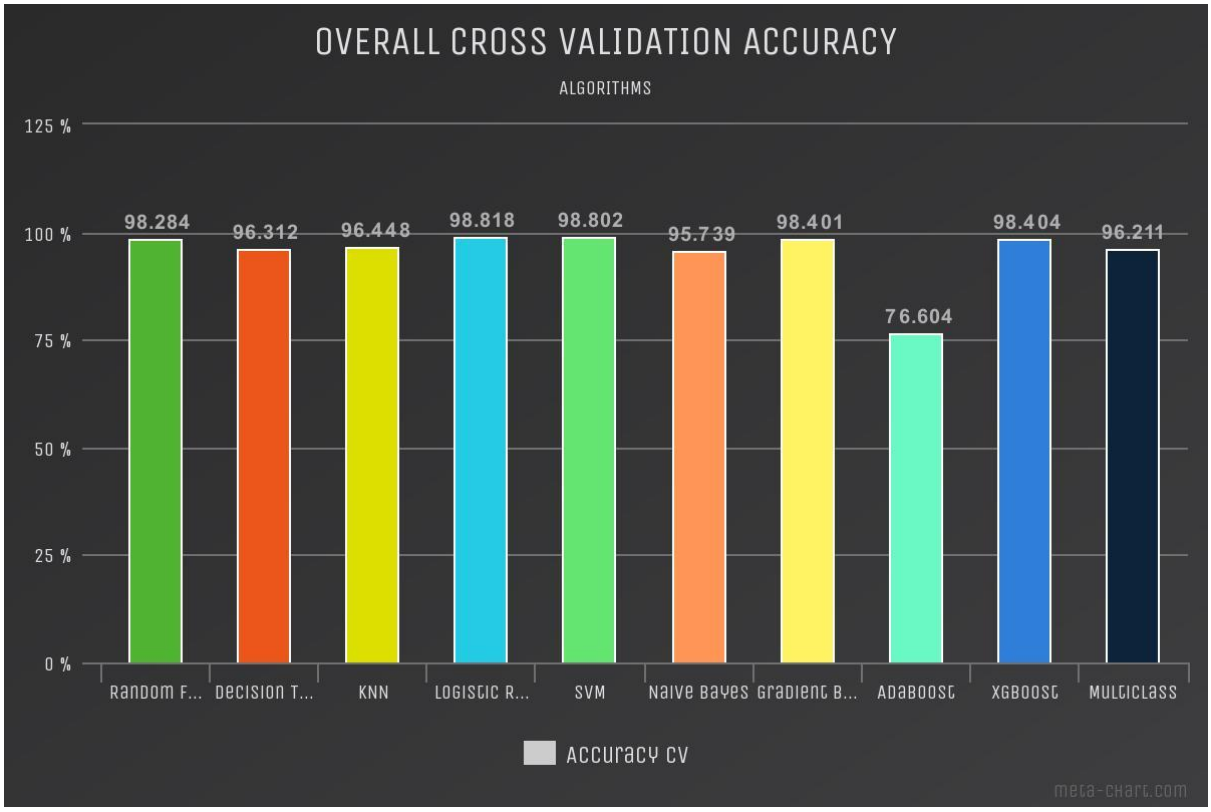

Figure: Cross Validation Accuracy (Overall)

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:

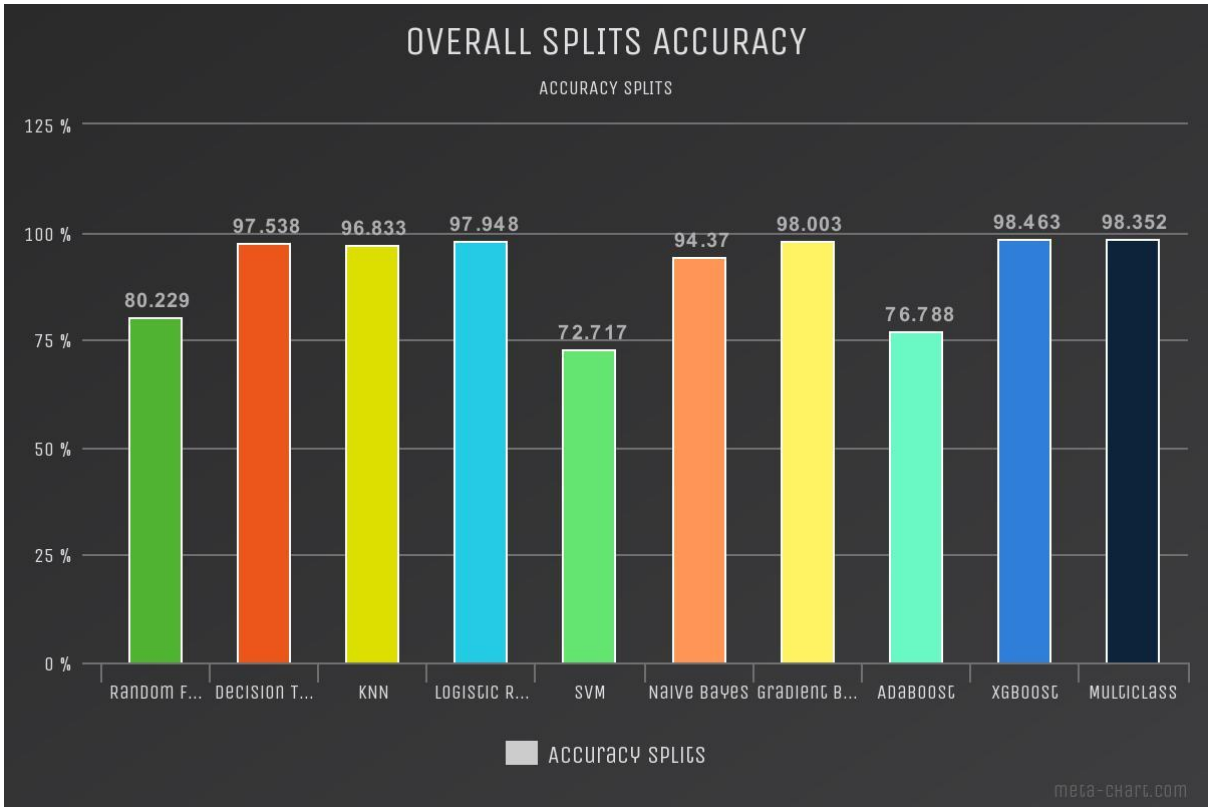

Figure: Splits Accuracy (Overall)

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:

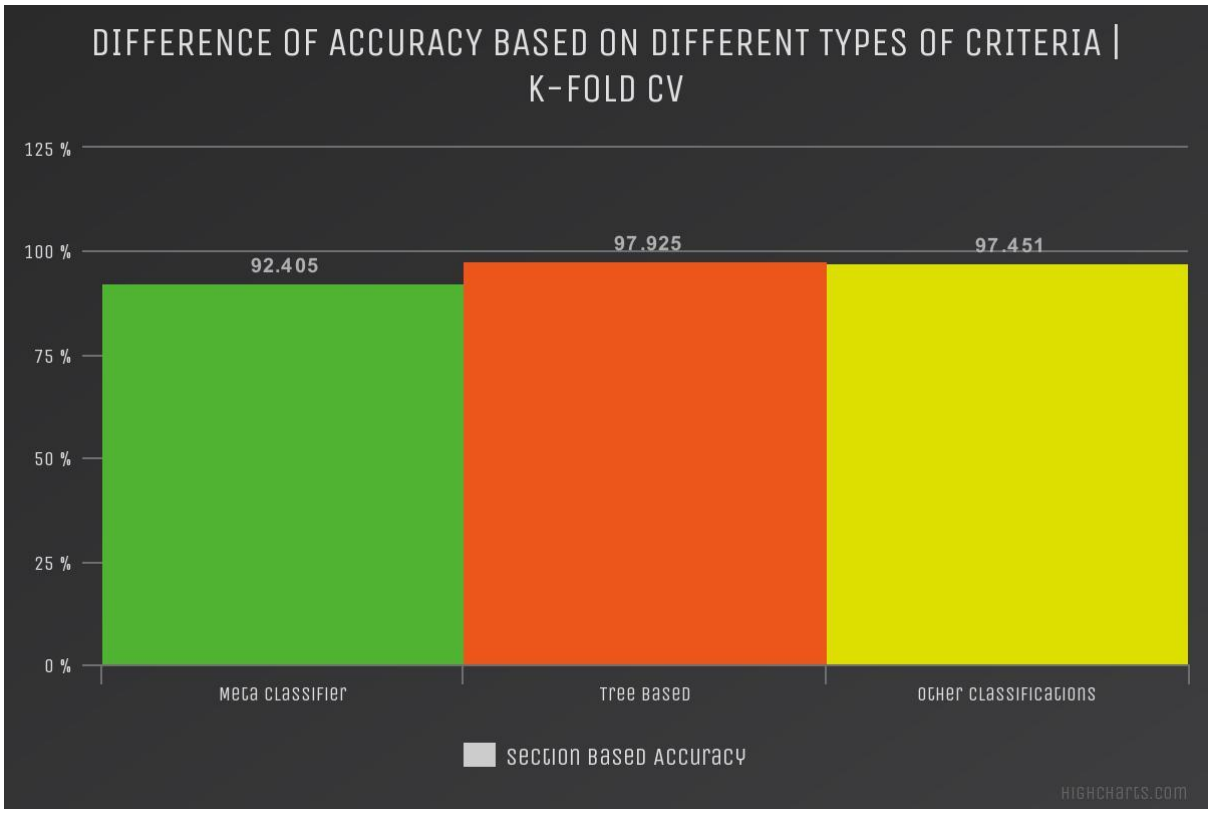

Figure: Accuracy Based on Different types of algorithms (Cross Validation)

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:

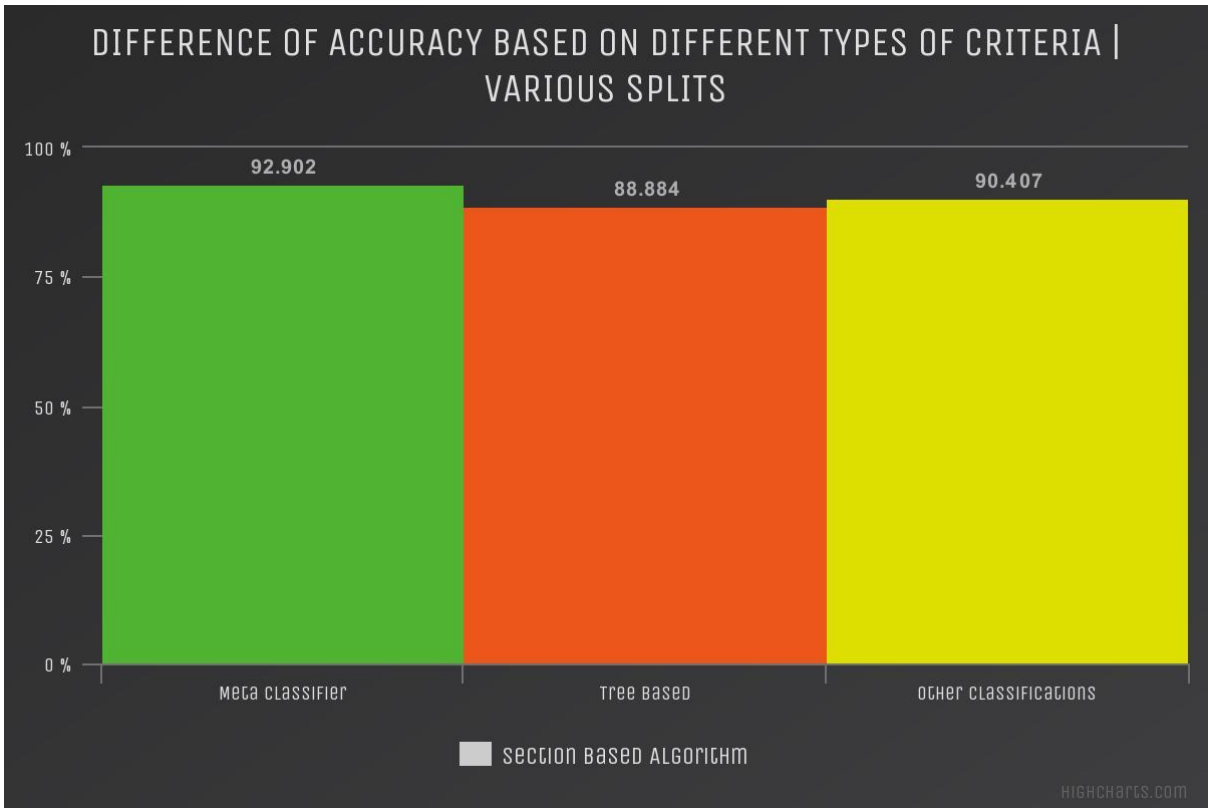

Figure: Accuracy Based on Different types of algorithm (Splits)

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:

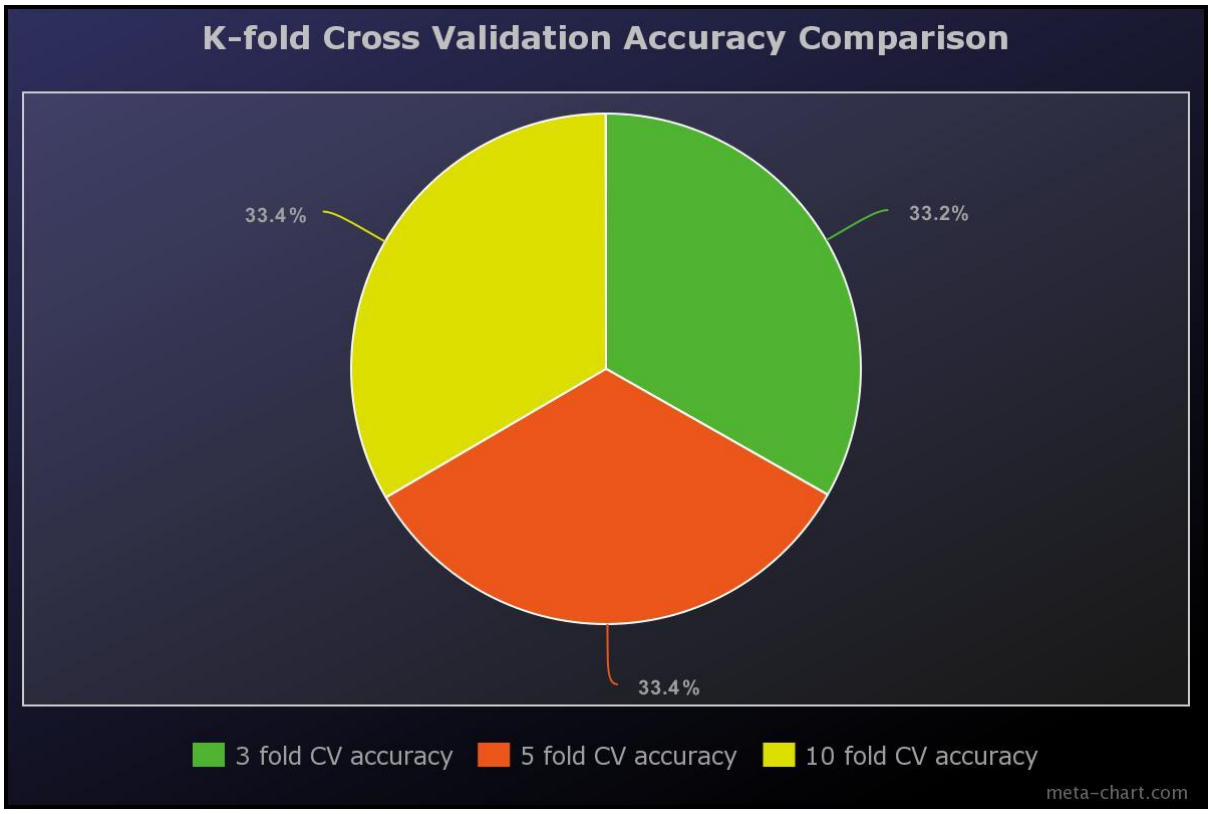

Figure: Accuracy Comparison (K-fold CV)

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:

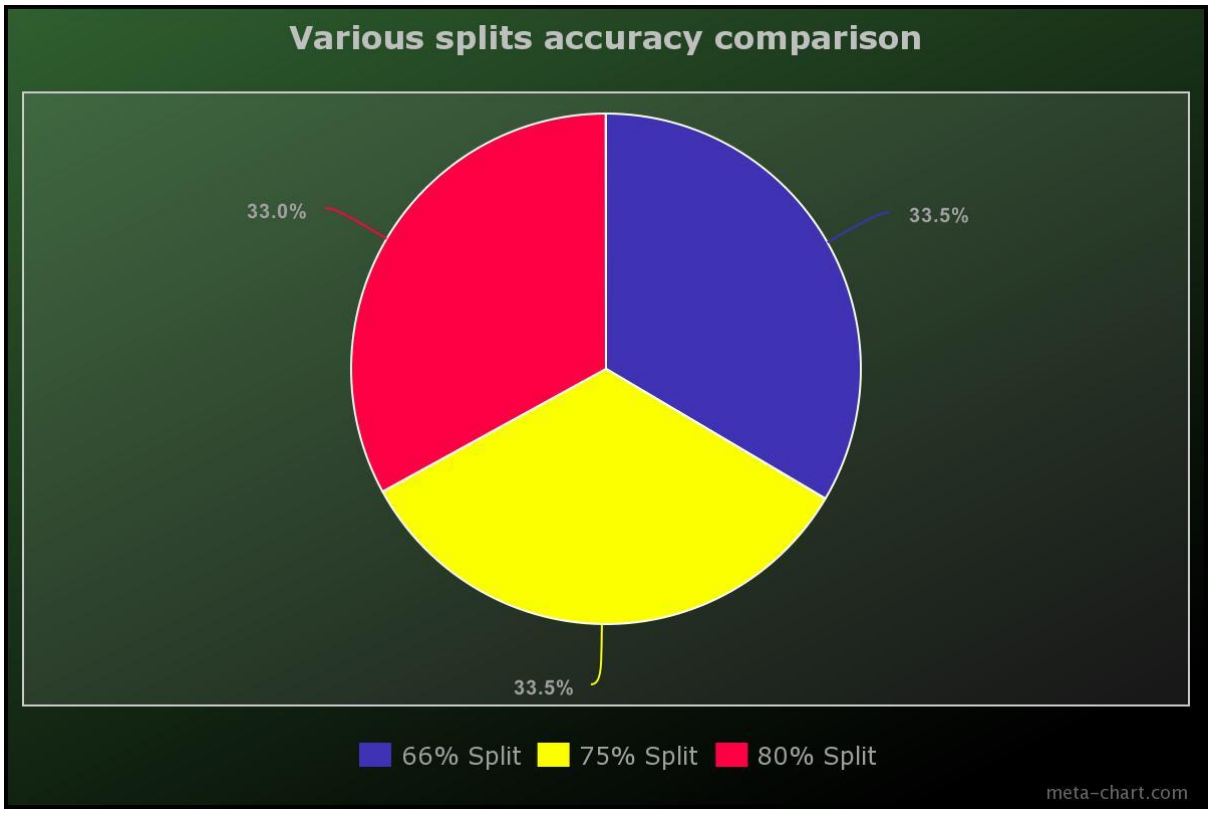

Figure: Accuracy comparison (splits)

Cross validation VS various splits:

In this part, we compare the 2 method 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 perform slightly well than various splits.

The chart is given below:

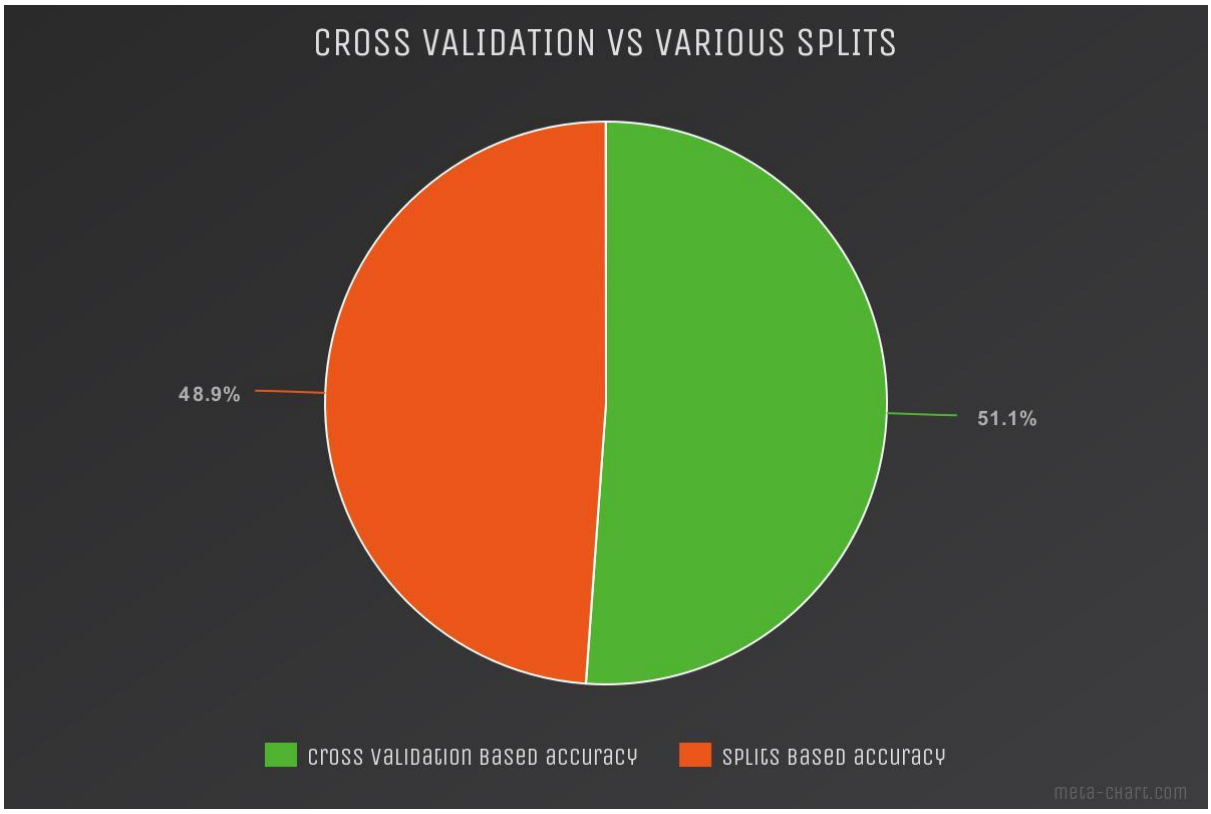

Figure: Cross Validation VS Splits

Chapter – 6

Reference

Chapter -7

Conclusion