

Extracting Meta Knowledge from Machine Learning and Data Mining Algorithms



**Meshu Deb Nath
Mehbuba Tabannur
Md. Mahamudul Hasan Tonoy
Susmita Saha**

**Department of Computer Science and Engineering
United International University**

*A thesis submitted for the degree of
BSc in Computer Science & Engineering*

May 2018

Declaration

The work presented in this thesis is the outcome of an investigation performed by Meshu Deb Nath, Mehbuba Tabannur, Md. Mahamudul Hasan Tonoy and Susmita Saha under the supervision of Dr. Dewan Md. Farid, Associate Professor, Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh.

We declare that this thesis titled, "Extracting Meta Knowledge from Machine Learning and Data Mining Algorithm" and the work presented in it are our own. We confirm that:

- This work was done wholly or mainly while in candidature for a BSc degree at United International University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.

Signed: Meshu

Date: May 09, 2018

(Meshu Deb Nath)

(ID: 011 132 058)

Signed: Mehbuba

Date: May 09, 2018

(Mehbuba Tabannur)

(ID: 011 132 034)

Signed: Mahamudul Hasan

Date: May 09, 2018

(Md Mahamudul Hasan Tonoy)

(ID: 011 132 031)

Signed: Susmita Saha

Date: May 09, 2018

(Susmita Saha)

(ID: 011 132 057)

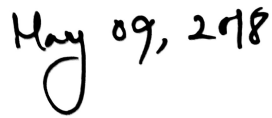
Certificate

I do hereby declare that the research works embodied in this thesis entitled “Extracting Meta Knowledge from Machine Learning and Data Mining Algorithms” is the outcome of an original work carried out by Meshu Deb Nath, Mehbuba tabannur, Md. Mahamudul Hasan Tonoy and Susmita Saha under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of BSc in Computer Science and Engineering.



Signed: _____



Date: _____

(Dr. Dewan Md. Farid)
Associate Professor,
Department of Computer Science and Engineering,
United International University,
Dhaka-1209, Bangladesh.

Abstract

Extracting meta-knowledge from real-world meta-data is a challenging task, which is a fundamental conceptual instrument for knowledge engineering and knowledge management. Meta-knowledge is a knowledge that learns to employ meta-learning in the field of machine learning and data mining. Recently, meta-information are using for decision making in many real-life machine learning for data mining applications. In this thesis, we have proposed a new method for extracting meta-knowledge applying machine learning algorithms. We have considered 10 most popular supervised learning algorithms (e.g. naive Bayes Classifier, Decision Tree, RandomForest, RandomTree, Bagging, Boosting, OneR, PART, Support Vector Machine, k-Nearest Neighbor) and 51 benchmark datasets from UC Irvine Machine Learning Repository. Initially, we have applied the learning algorithms on the datasets and engender the experimental results based on accuracy with 10-fold cross-validation and test data (30% split). Then, we have produced meta-data from the experimental results. To generate meta-data we added several extra features (i.e. size of the dataset, area, dimension, attribute types, is missing data present, streaming data etc.). Finally, we have applied decision tree induction algorithm on the meta-datasets to get meta-knowledge. We have tested our proposed model on several test data. The experimental results show that our proposed model achieved about 98% accuracy.

This thesis is dedicated to our parents.

Acknowledgements

We would like to express our sincere gratitude to our supervisor Dr. Dewan Md. Farid for the continuous support of research, for continuous guidance, for his patience, motivation and knowledge. His friendly attitude and inspiration throughout the research kept us motivated and optimistic.

We are thankful to Dr. Swakkhar Shatabda for his useful advice in developing our ideas. Additionally, we thank Prof. Dr. Chowdhury Mofizur Rahman and Dr. Salekul Islam for spending time in giving us a foundation on which to build our Research. Also, our sincere gratitude goes to United International University for providing us with a wonderful environment of academics.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Objectives of Thesis	2
1.3	Thesis Contribution	3
1.4	Organization of the Thesis	3
2	Learning Algorithm	4
2.1	Algorithm Descriptions	5
2.1.1	Naive Bayes	5
2.1.2	Random Tree and Random Forest	6
2.1.3	Decision Tree	6
2.1.4	Bagging	7
2.1.5	Boosting	7
2.1.6	OneR	8
2.1.7	Support Vector Machine	9
2.1.8	K-Nearest Neighbor	9
3	Dataset	11
3.1	Dataset Information	12
4	Meta Knowledge	19
4.1	Model Evaluation	19
4.2	Meta Data	20
5	Experimental Analysis	21
5.1	Experimental Results	21
5.1.1	Evaluation Results	21
5.1.2	Evaluation of Meta data	28
5.2	Extracting Meta-Knowledge	38
5.3	Performance Analysis	46

6	Conclusions and Future Work	50
6.1	Conclusions	50
6.2	Future Work	50
	Bibliography	51
A	Sample code	54

Chapter 1

Introduction

1.1 Motivation

Meta learning is a new interesting field in the machine learning community. The term meta learning first appeared in the area of educational psychology. One of the most cited researchers in this field, John Biggs, described meta learning as being aware of and taking control of ones own learning [1]. Hence, meta learning is defined as an understanding and adaptation of learning itself on a higher level than entirely acquiring subject knowledge. In that way, a person having the ability of meta learning is able to evaluate his or her learning approach and improve it according to the requirements of a specific task [1],[2].

Meta learning as used in a machine learning context has many similarities to this description. Meta learning is the study of principled methods that exploit meta knowledge to obtain efficient models and solutions by adapting machine learning and data mining processes [3],[4]. In the last 20 years, machine learning research was faced with an increasing number of algorithms but a lack of guidelines to select the right algorithm according to the nature of the problem. Meta learning can provide a valuable help avoiding long trial and error procedures for algorithm selection, and brute force searches for proper parametrisation [1]. Meta learning becomes meta learning by looking at different problems, domains, tasks or contexts or simply past experience.

Meta learning differs from the traditional view of learning also known as base learning in the scope of the level of adaptation. Learning at the base level is focused on accumulating experience on a specific learning task, whereas learning at the meta level is concerned with accumulating experience on the performance of multiple application of a learning system. It studies how a learning system can become more efficient by experience. The problem can be cast as that of determining the right bias for each task. Meta learning covers both declarative and procedural bias. Declarative bias specifies the representation of the space of hypotheses, Procedural bias imposes constraints on the ordering of the inductive hypotheses [5]. Searching through the bias space causes a meta learning algorithm to involve in a time-consuming process. An important aim

of meta learning is to exploit meta knowledge to make the search above the bias space flexible [3].

Meta knowledge is knowledge about a preselected knowledge. A very common form of meta knowledge is the performance of algorithms in certain problem domains, which is to be linked with characteristics of the task [1]. Meta knowledge contained different forms of information. It may include, for instance, a set of learning algorithms that have shown a good performance on datasets similar to the one under analysis, algorithms and datasets. Meta knowledge encompasses information useful to perform dynamic bias selection. It also includes functions and algorithms that helps to generate new useful information. It does not generally completely eliminate the need for search, but rather provides a more effective way of searching through the space of alternatives [3].

1.2 Objectives of Thesis

Consider the problem of selecting or recommending a suitable subset of machine learning algorithm for an assigned task. The problem can be defined as a search problem, where the aim is to identify the set of learning algorithms with the best performance and the search space includes the individual machine learning algorithms. Also our aim is to learn how to extract and exploit meta knowledge to benefit from previous experience.

A general framework for generating meta knowledge is illustrated in Figure 1.1. According to the framework, the process can be divided into two phases. In the first phase, the aim is to evaluate the machine learning algorithms using datasets. The second phase uses the output of the first phase to generate meta data and then generates meta knowledge by evaluating the meta data using decision tree algorithms. There are other objectives of our thesis for example

- Finding similar context of data set.
- Finding similarities between data set type and size.

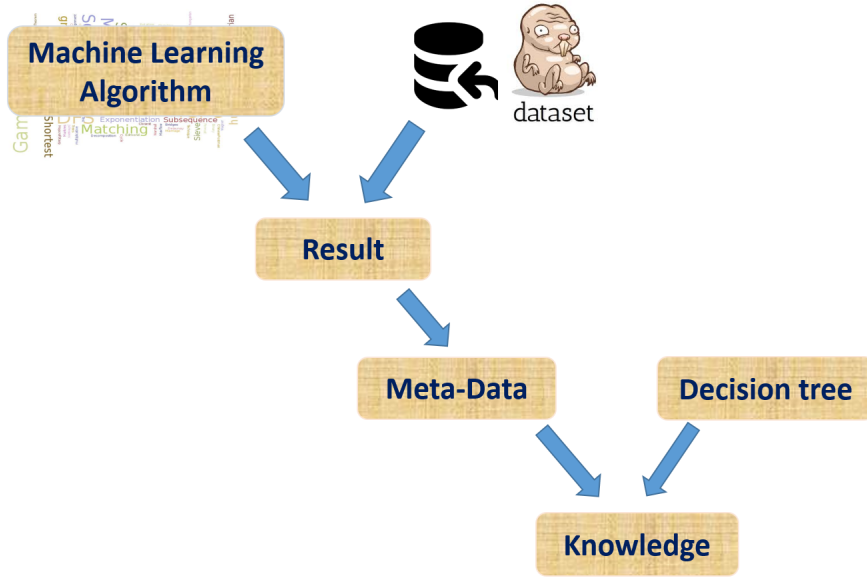


FIGURE 1.1: A general framework for generating meta knowledge.

1.3 Thesis Contribution

To generate meta data we used 51 datasets and 10 machine learning algorithms based on the popularity. First of all, we evaluated the machine learning algorithms using the datasets. In particular, we generated 1020 outputs using ten-fold cross-validation and test data (30% split). Then, the output of this evaluation helps to generate meta data where meta data is formed by characteristics of the data in the learning problem, and characteristics of the learning algorithm (type, parameter settings, performance measures) [6]. After that, we extracted meta knowledge from the meta data using decision tree algorithm. A meta knowledge can predict the learning algorithms best suited for the new problem [3].

1.4 Organization of the Thesis

The thesis is organized as follows:

In chapter 2, we present top 10 popular machine learning algorithm.

In chapter 3, we describe the datasets we used to evaluate our method.

In chapter 4, we present meta data and meta knowledge.

In chapter 5, we show the output tables of model evaluations and the extracting meta knowledge.

Then, we present the performance analysis of the meta knowledge.

In chapter 6, we describe conclusion part including future work.

Chapter 2

Learning Algorithm

The field of machine learning has matured a lot in the last decade and changed a lot in the last few years. Machine learning deals with analysis and creation of algorithms that facilitate the making of data-driven decisions from previously provided data [7]. Machine learning algorithms are typically classified into two categories: supervised learning and unsupervised learning.

Supervised Learning

Data mining is the process of finding hidden information and patterns from a huge database. Data mining algorithms have two major functions: classification and clustering. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data [8]. Supervised learning is the process of identifying new or unknown instances by classification algorithms based on a group of instances with known class membership. Supervised learning can be broadly classified into two subgroups: regression and classification. Regression is the problem of estimating or predicting a continuous quantity. Whereas, classification maps data into predefined groups or classes [9]. Classification deals with assigning observations into discrete categories, rather than estimating continuous quantities. Some popular examples of supervised machine learning algorithms are

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

Unsupervised Learning

Unsupervised learning consists of only input data and no corresponding input variables. Unsupervised learning creates the underlying structure or distribution in the data in order to learn more about the data. Unsupervised learning can be further grouped into clustering and association problems. A clustering problem refers to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour. Whereas, An association learning problem refers to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y [9]. Some popular examples of unsupervised machine learning algorithms are

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

2.1 Algorithm Descriptions

We applied machine learning algorithms from five different categories based on popularities. A tree diagram is illustrated in figure 2.1. From the tree-based algorithms, we applied decision tree, random forest, random tree. From ensemble learning algorithms, we applied bagging and boosting algorithms. From the rule-based algorithms, we applied oneR and Part. Furthermore, from the functional and lazy learning algorithms, we applied Support Vector Machine (SVM) and K-Nearest Neighbor (kNN) algorithm.

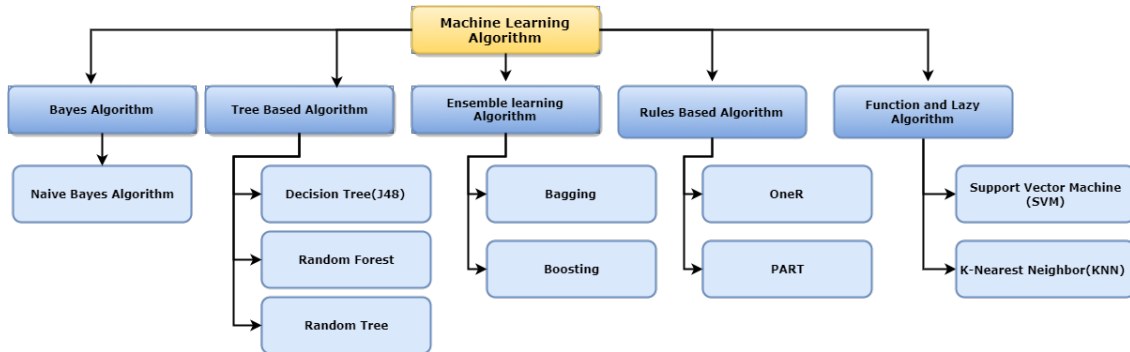


FIGURE 2.1: Tree diagram of algorithms used to evaluate datasets in this thesis.

The following section provides description and pseudocode of the algorithms.

2.1.1 Naive Bayes

A naive Bayes classifier is a simple probabilistic based method, which can predict the class membership probabilities [10]. This classifier can easily handle missing attribute values by simply omitting the corresponding probabilities for those attributes when calculating the likelihood of

membership for each class. And requires the class conditional independence, i.e. the effect of an attribute on a given class is independent of those of other attributes [11].

Algorithm 1 Naive Bayes classifier

Input: $D = \{x_1, x_2, \dots, x_n\}$ // Training data

Output: A Naive Bayes Model.

Method:

```

1: for each class,  $C_i \in D$ , do
2:   Find the prior probabilities,  $P(C_i)$ 
3: end for
4: for each attribute,  $A_i \in D$ , do
5:   for each attribute value,  $A_{ij} \in A_i$ , do
6:     Find the class conditional probabilities,  $P(A_{ij} \mid C_i)$ 
7:   end for
8: end for
9: for each instance,  $X_i \in D$ , do
10:  Find the posterior probability,  $P(C_i \mid X_i)$ 
11: end for
  
```

2.1.2 Random Tree and Random Forest

The Random Forest also called Random Decision Forest is an ensemble learning method for classification and regression, which is able to classify a large number of data. It develops various of decision trees based on a randomly selected subset of features at training time and outputting the class that is the ensemble of trees vote for the most popular class [12],[13].

Some random forests reported in the literature have consistently lower generalization error than others. For instance, random split selection does better than bagging. Breiman's introduction of random noise into the outputs also does better. But none of these three forests does as well as AdaBoost or other algorithms that work by adaptive re-weighting (arcing) of the training set [7],[14].

2.1.3 Decision Tree

Decision tree classification provides a rapid and useful solution for classifying instances in large datasets with a large number of variables. Decision tree can be constructed from a large volume of a dataset with many attributes because the tree size is independent of the dataset size. A decision tree has three main components: nodes, leaves, and edges [15].

The goal of a decision tree is to create a model that predicts the value of the target class for an unseen test instance based on several input instances. J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in the classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in the classification for that tuple [16].

Algorithm 2 Decision Tree

Input: $D = \{x_1, x_2, \dots, x_n\}$
Output: A decision tree, DT.

Method:

```

1:  $DT = \phi$ ;
2: find the root node with best splitting,  $A_j \in D$ ;
3:  $DT =$  create the root node;
4:  $DT =$  add arc to root node for each split predicate and label;
5: for each arc do
6:    $D_j$  created by applying splitting predicate to  $D$ ;
7:   if stopping point reached for this path then
8:      $DT' =$  create a leaf node and label it with  $cl$ ;
9:   else
10:     $DT' = DTBuild(D_j)$ ;
11:   end if
12:  $DT =$  add  $DT'$  to arc;
13: end for

```

2.1.4 Bagging

Bagging creates subsets from the original dataset using sampling with replacement method. It then derives a model for each subset using a learning scheme [17]. The bagging classifier shows better predictive accuracy compared to the individual base classifiers in the case of balanced classification tasks [18],[19].

Algorithm 3 Bagging Algorithm

Input: Training data, D , number of iterations, k , and a learning scheme.

Output: Ensemble model, M
Method:

```

1: for  $i=1$  to  $k$  do
2:   create bootstrap sample  $D_i$ , by sampling  $D$  with replacement;
3:   use  $D_i$ , and learning scheme to derive a model,  $M_i$ ;
4: end for

```

To use M to classify a new instance, x_{New} : Each M_i M^* classify x_{New} and return the majority vote;

2.1.5 Boosting

Boosting is an ensemble method of generating strong hypothesis through the combination of multiple weak classifiers. Boosting is similar to bagging in that it combines multiple base learners to obtain a result based on voting technique [20],[21]. The concept of adaptive boosting called AdaBoost algorithm was first introduced by Freund and Schapire in 1997 [35] that classify an example by voting the weighted predictions of a set of base classifiers, which are generated in a series of rounds [22]. However, its differences lie in that boosting assigns weight to instances according to how hard they are to classify, thus setting a high weight of instances that are difficult to classify. One base learner contributes to the weight used by the next base learner. Weights are

assigned to the base learners as well according to their predictive accuracy and this is taken into consideration when classifying a new test instance [23],[24],[25].

Algorithm 4 AdaBoost Algorithm

Input: Training data, D , number of iterations, k , and a learning scheme.

Output: Ensemble model, M **Method:**

```

1: initialize weight,  $x_i \in D$  to  $\frac{1}{d}$ ;
2: for  $i=1$  to  $k$  do
3:   sample  $D$  with replacement according to instance weight to obtain  $D_i$ ;
4:   use  $D_i$ , and learning scheme to derive a model,  $M_i$ ;
5:   compute  $\text{error}(M_i)$ ;
6:   if  $\text{error}(M_i) \geq 0.5$  then
7:     go back to step 3 and try again;
8:   end if
9:   for each correctly classied  $x_i \in D$  do
10:    multiply weight of  $x_i$  by  $\frac{\text{error}(M_i)}{1-\text{error}(M_i)}$ ;
11:   end for
12:   normalise weight of instances;
13: end for

```

To use M * to classify a new instance, x_{New} :

```

1: for  $i=1$  to  $k$  do
2:    $w_i = \log \frac{1-\text{error}(M_i)}{\text{error}(M_i)}$ ; // weight of the classifiers vote
3:    $c = M_i(x_{New})$ ; // class prediction by  $M_i$ 
4:   add  $w_i$  to weight for class  $c$ ;
5: end for
6: return class with largest weight;

```

2.1.6 OneR

OneR is rules based classification on the value of a single predictor. It generates one rule for each predictor in the data, then selects the rule with the smallest total error as its one rule. It is a simple, cheap method that often produces good rules with high accuracy for characterizing the structure in data. This classifier is often used as a baseline for the comparison between the other classification models, and as an indicator of the predictive power of attributes [26],[27].

Algorithm 5 OneR algorithm

Input: $D = \{x_1, x_2, \dots, x_n\}$ // Training data**Output:** OneR Model**Method:**

```
1: for each attribute,  $A_i \in D$ , do
2:   for each attribute value,  $A_{ij} \in A_i$ , do
3:     Make a classification rule:
4:       count how often each class appears
5:       find the most frequent class
6:       make the rule assign that class to this  $A_{ij}$ 
7:   end for
8:   Calculate the error rate of this attributes  $A_i$  rule.
9: end for
10: Choose the attribute  $A_i \in D$  with the smallest error rate.
```

2.1.7 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space with each feature being the value of particular coordinate. SVM are state-of-the-art large margin classifiers which have recently gained popularity within visual pattern recognition [28]. Training a SVM leads to a quadratic optimization problem with bound constraints and one linear equality constraint. Despite the fact that this type of problem is well understood there are many issues to be considered in designing an SVM learner. In particular for large learning tasks with many training examples off-the-shelf optimization techniques for general quadratic programs quickly become intractable in their memory and time requirements [29],[30].

2.1.8 K-Nearest Neighbor

The K-nearest neighbors (KNN) is a classification algorithm based on the use of distance measures [31]. K-nearest neighbors classifier can select an appropriate value of K based on cross-validation. k-NN is a simple classifier, which uses the distance measurement techniques that widely used in machine learning.

The k-NN algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct. The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance [26].

Algorithm 6 k-nearest-neighbor classifier

Input: $D = \{x_1, x_2, \dots, x_n\}$

Output: kNN classifier, kNN

Method:

- 1: Find $X \in D$ that identify the k nearest neighbours, regardless of class label, C_l .
 - 2: out of these instances, $X = x_1, x_2, \dots, x_k$, identify the number of instances, k_i , that belong to class c_l , $l = 1, 2, \dots, M$. Obviously, $\sum_i k_i = k$
 - 3: assign X_{test} to the class C_l with the maximum number of K_i of instances.
-

Chapter 3

Dataset

A machine-learning algorithm is a mathematical model that learns to find patterns in the input that is fed to it. One of the challenging problem is to feed the right data in the right format. Feeding the right data means collecting the data that correlates with the output we want to predict. Machine learning typically works with two data set such as training data set and testing data set [32]. In this thesis, we applied 51 dataset to evaluate the learning algorithm. We illustrated a general feature of the datasets used in this thesis in Figure 3.1.

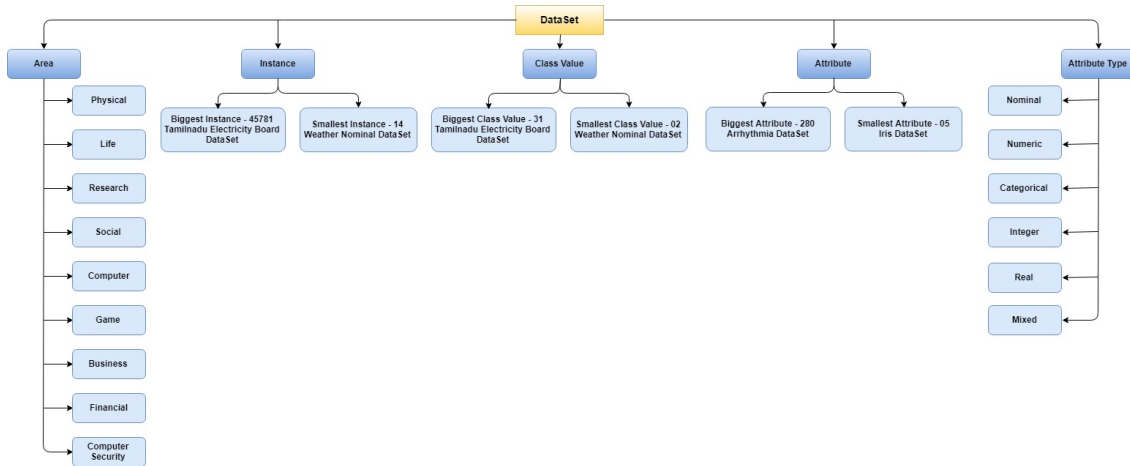


FIGURE 3.1: Feature of Dataset.

According to the Figure 3.1, datasets are collected from different area such as life, social, computer, business, physical etc. From all the collected datasets, highest instance number is 45781 and lowest instance number is 14. Also, highest class value is 31 and lowest class value is 2. Similarly, highest attribute number is 280 and lowest attribute number is 5.

3.1 Dataset Information

All the dataset used in this thesis are collected from UCI machine learning repository [33]. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. Table 3.1 describes the datasets used in our experimental analysis to evaluate the machine learning algorithms.

Iris Dataset:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Balance Scale Dataset:

This dataset was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance.

Liver Disorders Dataset:

The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the dataset constitutes the record of a single male individual.

Car Evaluation Dataset:

Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making.

Diabetes Dataset:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

GPS Trajectories Dataset:

The dataset is composed by two tables. The first table go tracks presents general attributes and each instance has one trajectory that is represented by the table go track tracks points.

Glass Identification Dataset:

Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis. BEAGLE is a product available through VRS Consulting.

Breast Cancer Dataset:

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature.

This dataset includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Tic-Tac-Toe Endgame Dataset :

This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where “x” is assumed to have played first. The target concept is “win for x” (i.e., true when “x” has one of 8 possible ways to create a “three-in-a-row”).

Page Blocks Classification Dataset:

The 5473 examples comes from 54 distinct documents. Each observation concerns one block. All attributes are numeric.

MAGIC Gamma Telescope Dataset:

The data are MC generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. Cherenkov gamma telescope observes high energy gamma rays, taking advantage of the radiation emitted by charged particles produced inside the electromagnetic showers initiated by the gammas, and developing in the atmosphere.

Wine Quality Dataset:

The dataset is related to red and white variants of the Portuguese “Vinho Verde” wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available. This dataset can be viewed as classification or regression tasks. The classes are ordered and not balanced. Outlier detection algorithms could be used to detect the few excellent or poor wines.

Vowels Dataset:

The data was collected for examining our newly developed classifier for multidimensional curves (multidimensional time series). Nine male speakers uttered two Japanese vowels successively. For each utterance, with the analysis parameters described below, we applied 12-degree linear

prediction analysis to it to obtain a discrete-time series with 12 LPC cepstrum coefficients. This means that one utterance by a speaker forms a time series whose length is in the range 7-29 and each point of a time series is of 12 features (12 coefficients).

EEG Eye State:

All data is from one continuous EEG measurement with the Emotiv EEG Neuroheadset. The duration of the measurement was 117 seconds. The eye state was detected via a camera during the EEG measurement and added later manually to the file after analysing the video frames. '1' indicates the eye-closed and '0' the eye-open state. All values are in chronological order with the first measured value at the top of the data.

Credit Approval Dataset:

This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

Vote Dataset:

This dataset includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

Thoracic Surgery Data Dataset:

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007 to 2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

Pen-Based Recognition of Handwritten Digits Dataset:

This dataset has a digit database by collecting 250 samples from 44 writers. The samples written by 30 writers are used for training, cross-validation and writer dependent testing, and the digits

written by the other 14 are used for writer independent testing.

Letter Recognition Dataset:

This dataset has 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes which were then scaled to fit into a range 0 to 15.

Statlog (Vehicle Silhouettes) Dataset:

The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTs, which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness.

Seismic-bumps Dataset:

Mining activity was and is always connected with the occurrence of dangers which are commonly called mining hazards. A special case of such threat is a seismic hazard which frequently occurs in many underground mines. Seismic hazard is the hardest detectable and predictable of natural hazards and in this respect it is comparable to an earthquake. More and more advanced seismic and seismoacoustic monitoring systems allow a better understanding rock mass processes and definition of seismic hazard prediction methods. Accuracy of so far created methods is however far from perfect.

Mushroom Dataset:

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended.

HIGGS Dataset:

The data has been produced using Monte Carlo simulations. The first 21 features are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes.

Flags Dataset:

This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country

from its size and the colours in its flag. 10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.

Ionosphere Dataset:

This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts.

Soybean Dataset:

There are 19 classes, only the first 15 of which have been used in prior work. The folklore seems to be that the last four classes are unjustified by the data since they have so few examples. There are 35 categorical attributes, some nominal and some ordered. The value “dna” means does not apply. The values for attributes are encoded numerically, with the first value encoded as ”0”, the second as “1,” and so forth. An unknown value is encoded as ”?”.

Waveform Dataset:

This dataset has 3 classes of waves, 40 attributes, and noise values. The latter 19 attributes are all noise attributes with mean 0 and variance 1.

KDD Cup 1999 Dataset:

This is the dataset used for the third international knowledge discovery and data mining tools competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between “bad” connections, called intrusions or attacks, and “good” normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Spambase Dataset:

This dataset collect spam e-mails from postmaster and individuals who had filed spam. Personal emails let to collect non-spam e-mails from filed work. Hence ‘george’ and ‘650’ area code are indicators of non-spam. When generating a spam filter these are useful information.

Optical Recognition of Handwritten Digits Dataset:

This dataset extracted normalized bitmaps of handwriting digits from a pre-printed form. Over 43 people contributed to the training and testing set. 32x32 bitmaps are divided into non overlapping blocks of 4x4. This generates an input matrix of 8x8 where each element is an integer in the range 0 to 16.

Polish companies bankruptcy Datasets:

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS), which is a database containing information on emerging markets around the world.

Multiple Features Datasets:

This dataset consists of features of handwritten numerals 0 to 9 extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images.

Arrhythmia Datasets:

This database contains 279 attributes, 206 of which are linear valued and the rest are nominal. The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones.

TABLE 3.1: Dataset Description.

No	Dataset	Instance	Attribute	Attribute Type	Class value
1	Weather nominal	14	5	Nominal	2
2	Weather numeric	14	5	Numeric	2
3	Contact-lenses	24	5	Nominal	3
4	Iris	150	5	Numeric	3
5	Balance Scale	625	5	Categorical	3
6	Tamilnadu Electricity Board	45781	5	Integer, Real	31
7	Car Evaluation	1728	6	Categorical	4
8	Liver Disorders	345	7	Categorical, Integer, Real	2
9	Diabetes	768	9	Numeric	2
10	Tic-Tac-Toe Endgame	958	9	Categorical	2
11	Glass	214	10	Numeric	7
12	Breast-cancer	286	10	Nominal	2
13	Page Blocks Classification	5473	10	Integer, Real	5
14	Magic Gamma Telescope	19020	11	Real	2
15	Vowel Data	990	12	Real	11
16	Wine Quality	1599	12	Real	6
17	GPS Trajectories	163	15	Real	22
18	Credit Approval	690	15	Categorical, Integer, Real	2
19	EEG Eye State	14980	15	Integer, Real	2
20	Pen-Based Recognition	10992	16	Integer	10
21	Letter Recognition	20000	16	Integer	26
22	Labor	57	17	Numeric	2
23	Vote	435	17	Nominal	2
24	Thoracic Surgery	470	17	Integer, Real	2
25	Statlog (Vehicle Silhouettes)	846	18	Integer	4
26	Seismic-bumps	2584	19	Real	2
27	Segment-test	810	20	Numeric	7
28	Statlog (German Credit Data)	1000	20	Categorical, Integer	2
29	Segment-challenge	1500	20	Numeric	7
30	credit-g	1000	21	Nominal	2
31	mushroom	8124	22	Categorical	2
32	Parkinsons	1040	23	Real	2
33	Chronic Kidney Disease	400	25	Real	2
34	HIGGS	20000	29	Integer	2
35	Flags	194	30	Categorical, Integer	8
36	Hypothyroid	3772	30	Numeric	4
37	Phishing Websites	11055	31	Integer	2
38	Student alcohol consumption	1044	32	Integer	18
39	Unbalanced	856	33	Numeric	2
40	Alcohol consumption	1043	33	Integer	18
41	Ionosphere	351	35	Numeric	2
42	Soybean	683	36	Nominal	19
43	Annealing	898	38	Categorical, Integer, Real	6
44	Waveform	5000	40	Real	3
45	KDD Cup 1999	10001	42	Categorical, Integer	23
46	Spambase	4601	57	Integer, Real	2
47	Optical Recognition	5620	64	Integer	10
48	Polish companies bankruptcy	10503	64	Real	2
49	Supermarket	4627	217	Numeric	2
50	Multiple Features	2000	241	Integer, Real	10
51	Arrhythmia	452	280	Categorical, Integer, Real	16

Chapter 4

Meta Knowledge

In this chapter, we describe how to gain meta data from previous knowledge. Meta data is a set of data that describes and gives information about other data. The idea underlying meta-learning is to identify a meta-knowledge apparatus that could be useful in supporting the search for a suitable bias when applying base-learner to solve a specific task [6].

4.1 Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. The data used for training is not acceptable in data mining to evaluate model performance because it can create overoptimistic and overfitted models. There are two methods of evaluating models in data mining, Hold-Out and Cross-Validation. Both methods use a test set to evaluate model performance to avoid overfitting [34].

Hold-Out

In this method, the mostly large dataset is randomly divided into three subsets: training set, validation set and test set.

Cross-validation

When data is limited, we use k-fold cross-validation to achieve an unbiased estimate of the model performance. When data is limited, k-fold cross-validation is being used to achieve an unbiased estimate of the model performance. In k-fold cross-validation, the data is divided into k subsets of equal size.

The problem of predictive modelling is to create models that have good performance making predictions on new unseen data. Therefore it is important to use strong techniques to train and evaluate models on available training data. To evaluate the machine learning algorithms, we choose percentage split and ten-fold cross-validation model evaluation techniques from the Weka machine learning workbench.

4.2 Meta Data

Meta data means data about data. Meta data is defined as the data providing information about one or more aspects of the data [6],[1]. A meta data consists of several features such as properties of data set, knowledge from previous result etc.

In this thesis, a meta data has several features. For example, it has a feature defined as “size of dataset” where it has the instance number of the dataset. Instead of collecting raw instance number as a size of the dataset, we categorized instance number as small, medium, large, big-data. When instance number is less than 500, size of the dataset is defined as “small”. Similarly, if instance number is greater than 500 but less than 1000, size of the dataset is defined as “Medium”. When instance number is greater than 1000 but less than 10000 is defined as “Large”. Else if instance number is greater than 10000, size of the dataset is defined as “Big-data”. “Area” is another feature of a meta data. In this thesis, most of the area is physical, computer science, computer security, life, business, financial, social, game, research.

Meta data has another feature named as “Streaming Data” where it has a Boolean information: yes or no according to the nature of the dataset source. Also, a meta data has “Dimension” feature based on the attribute value of datasets, this feature has three values such as Low, Medium, High. Furthermore, meta data has a feature named as “Attribute type” where it has the attribute type information of dataset. Some attribute type is numeric, nominal, mixed etc.

A meta data also has a class feature where the value generated according to the performance of the algorithm. We compared the accuracy of the previous experimental results with a standard accuracy. We generated the standard accuracy by evaluating datasets using a baseline classifier “zeroR”. Therefore, we consider three different rules to generate the class value, if the classifier accuracy is greater than 90% and also greater than standard accuracy then the class is “Good”. Else if the classifier accuracy is greater than standard accuracy but less than 90% then the class is “Average”. Else if the classifier accuracy is less than standard accuracy then the class is “Bad”. Thus, we have generated Meta-data for the machine learning algorithms.

Chapter 5

Experimental Analysis

In this chapter, we present experimental results, extracting meta-knowledge and performance analysis of the meta knowledge.

5.1 Experimental Results

Before generating the meta knowledge first, we need to evaluate the machine learning algorithms. To evaluate the performance of the algorithms we used ten-fold cross-validation and test data (30% split).

5.1.1 Evaluation Results

We described 10 machine learning algorithms and 51 datasets in chapter 3 and chapter 4. Here, we present the experimental results of the machine learning algorithms by evaluating datasets. We created output tables using the performance of the machine learning algorithms. We sorted the result tables by attribute number of the related datasets. We illustrated a portion of an output table in Table 5.1(a), where we evaluated the algorithms by ten-fold cross-validation. Table 5.1(a) represented 1 to 15 instances. Thus, Table 5.1(b) represented 16 to 31 instances and Table 5.1(c) represented 32 to 51 instances of this output table.

In Table 5.2(a), 5.2(b), 5.2(c) we illustrated another output table generated by percentage splitting evaluation method (30% test data).

Table 5.1(a): Output table by ten-fold cross-validation (1-15 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	Random Forest	Random Tree	SMO
1	Weather Nominal	64.29	57.14	57.14	42.86	50.00	50.00	50.00	71.43	57.14	57.14
2	Weather Numeric	50.00	57.14	92.86	57.14	50.00	42.86	35.71	42.86	35.71	64.29
3	Contact-Lenses	70.83	66.67	75.00	83.33	70.83	70.83	83.33	70.83	75.00	66.67
4	Iris	96.00	94.00	96.67	94.00	96.00	92.00	94.00	94.00	94.67	96.00
5	Balance Scale	72.32	83.36	89.44	76.64	90.40	56.32	83.52	81.44	77.28	87.68
6	Tamilnadu Electricity	9.54	65.82	65.31	65.62	65.08	65.62	65.62	0.00	65.49	65.31
7	Liver Disorders	66.09	69.57	64.35	68.70	55.36	55.07	63.77	73.04	67.82	58.26
8	Car Evaluation	70.66	90.35	71.24	90.93	88.42	70.66	94.59	93.82	85.52	93.63
9	Diabetes	75.39	75.91	75.00	72.01	75.78	71.48	74.74	75.91	67.84	76.95
10	Gps Trajectories	51.25	53.75	52.50	58.75	58.75	50.00	58.75	61.25	58.75	52.50
11	Glass	44.86	71.96	62.15	64.49	47.66	56.54	69.63	80.37	69.16	56.54
12	Breast-Cancer	71.68	67.13	73.43	73.43	72.73	65.39	66.43	68.53	62.94	70.28
13	Tic-Tac-Toe Endgame	72.55	91.54	75.47	84.55	69.62	69.94	94.26	96.35	79.23	98.33
14	Page Block	93.30	97.50	89.46	97.44	82.16	94.34	97.02	97.50	96.41	92.08
15	Magic Gamma Telescope	78.09	86.96	78.38	85.08	72.66	70.80	85.20	88.01	81.60	79.14

Table 5.1(b): Output table by ten-fold cross-validation (16-31 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	Random Forest	Random Tree	SMO
16	Wine Quality	93.42	65.17	61.04	61.48	55.03	54.66	61.35	70.11	62.41	58.35
17	Vowel Data	15.49	98.59	33.00	77.78	58.59	29.97	68.01	93.60	81.14	61.28
18	Eeg Eye State	68.60	88.41	65.23	82.31	45.17	63.53	63.53	92.52	81.98	55.61
19	Credit Approval	61.63	52.17	85.65	85.22	84.93	85.51	83.33	75.51	62.61	86.67
20	Labor	92.98	89.47	91.23	82.46	94.74	73.68	80.70	91.23	84.21	94.74
21	Vote	95.63	95.86	91.26	96.32	90.12	95.63	96.55	96.32	94.48	96.09
22	Thoracic Surgery Data	84.04	83.83	85.11	84.47	78.51	83.40	79.15	83.62	74.89	84.89
23	Pen-Based Recognition	20.38	97.91	84.89	96.56	85.75	38.95	96.83	99.13	95.72	97.96
24	Letter Recognition	7.09	90.05	57.91	87.98	64.12	17.24	88.76	96.51	86.16	82.34
25	Statlog (Vehicle Silhouet	85.65	72.10	64.66	72.46	44.80	51.89	71.51	76.00	70.92	74.35
26	Seismic-Bumps	19.50	92.45	93.42	93.42	82.28	28.10	93.42	0.00	91.95	91.83
27	Segment-Test	28.64	94.57	85.80	93.33	87.28	60.99	92.35	95.80	91.36	91.98
28	Segment-Challenge	30.40	96.53	85.47	95.80	81.33	64.53	95.60	97.93	95.07	91.67
29	Statlog (Credit Data)	72.67	72.33	72.67	73.67	73.33	72.67	73.67	75.67	70.33	75.00
30	Credit-G	70.90	72.70	72.30	70.90	75.50	66.40	69.40	75.70	66.10	75.10
31	Mushroom	95.45	100.00	89.66	100.00	95.40	98.69	100.00	100.00	99.96	100.00

Table 5.1(c): Output table by ten-fold cross-validation (32-51 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	Random Forest	Random Tree	SMO
32	Chronic Kidney Disease	99.00	98.75	85.50	99.00	95.00	92.00	98.50	100.00	95.50	97.75
33	Parkinsons	51.03	67.31	65.29	62.79	59.04	56.44	59.23	70.10	63.08	63.08
34	Higgs	66.40	69.47	54.67	65.17	59.23	58.35	65.17	71.57	60.70	63.60
35	Flags	99.98	35.57	66.49	59.28	55.15	4.64	57.73	61.34	42.78	60.82
36	Hypothyroid	92.76	99.47	92.29	99.60	95.39	96.00	99.52	99.39	97.64	93.58
37	Phishing Websites	92.58	96.20	90.76	95.88	92.98	88.89	96.76	97.26	96.37	93.80
38	Student Alcohol Consumption	39.95	45.45	14.18	39.45	42.68	46.23	34.82	37.29	30.05	22.65
39	Unbalanced	98.60	98.60	98.60	98.60	89.60	98.60	97.78	98.60	97.55	98.60
40	Alcohol Consumption	20.13	46.88	28.86	39.41	39.02	49.86	40.17	45.06	30.97	40.27
41	Ionosphere	91.17	92.59	84.90	91.17	81.77	80.63	90.60	93.73	90.60	87.46
42	Soybean	27.97	84.63	68.67	91.65	92.39	39.97	91.22	92.97	83.90	92.83
43	Annealing	84.01	99.22	89.59	98.88	87.73	84.01	98.51	98.88	97.40	96.65
44	Waveform	66.64	81.52	84.96	75.08	80.00	53.72	77.42	85.06	73.08	86.68
45	Kdd Cup 1999	99.87	99.93	99.76	99.89	99.82	99.90	99.94	99.97	99.95	99.98
46	Spambase	90.07	94.00	83.26	92.98	79.29	78.33	93.59	95.50	90.94	90.42
47	Optical Recognition	19.75	94.23	93.86	90.69	91.33	27.03	92.54	98.15	85.94	98.33
48	Bankruptcy	95.24	95.42	95.29	95.13	9.89	95.19	95.22	95.02	92.05	95.28
49	Supermarket	75.38	63.71	37.07	63.71	63.71	66.74	63.71	63.71	63.71	63.71
50	Multiple Features	27.74	84.00	89.95	78.65	93.30	23.10	82.00	96.50	65.35	97.55
51	Arrhythmia	55.53	72.12	55.53	64.38	62.39	60.40	63.94	67.26	51.11	70.13

Table 5.2(a): Output table by test (30%) (1-15 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	RandomForest	RandomTree	SMO
1	Weather Nominal	50.00	50.00	25.00	25.00	50.00	25.00	25.00	50.00	25.00	50.00
2	Weather Numeric	25.00	50.00	50.00	25.00	50.00	50.00	25.00	50.00	25.00	50.00
3	Contact-Lenses	28.57	28.57	42.86	28.57	28.57	28.57	42.86	42.86	28.57	28.57
4	Iris	95.56	95.56	95.56	95.56	95.56	95.56	95.56	95.56	93.33	95.56
5	Balance Scale	68.45	80.21	85.56	77.01	86.63	57.75	79.68	79.14	77.54	85.56
6	Tamilnadu Electricity	9.31	65.41	64.58	64.58	64.96	64.58	64.58	0.00	65.99	64.69
7	Liver Disorders	75.73	67.96	64.08	71.84	55.34	55.34	64.08	73.79	66.02	63.11
8	Car Evaluation	70.02	96.70	77.37	96.30	87.09	70.02	98.67	100.00	100.00	95.14
9	Diabetes	79.57	77.39	77.39	76.52	76.96	74.35	72.61	76.52	70.00	79.13
10	Gps Trajectories	52.17	52.17	47.83	65.22	56.52	52.17	65.22	56.52	56.52	56.52
11	Glass	42.19	65.63	59.38	60.94	42.19	54.69	60.94	78.13	71.88	67.19
12	Breast-Cancer	73.26	66.28	66.28	63.95	67.44	63.95	62.79	69.77	59.30	72.09
13	Tic-Tac-Toe Endgame	71.78	84.67	74.91	80.84	70.73	70.38	94.43	90.59	74.56	98.95
14	Page Blocks Classification	93.13	98.19	97.50	98.48	90.10	94.15	98.54	99.74	99.74	93.00
15	Magic Gamma Telescope	81.11	87.35	77.57	85.51	72.84	70.75	85.16	88.05	81.11	79.20

Table 5.2(b): Output table by test (30%) (16-31 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	RandomForest	RandomTree	SMO
16	Wine Quality	57.72	61.03	61.04	55.63	51.04	55.63	59.17	66.67	56.88	59.17
17	Vowel Data	17.78	76.09	10.30	97.68	71.21	43.94	96.77	100.00	100.00	85.46
18	Eeg Eye State	71.00	96.76	65.87	97.22	44.87	63.81	63.81	100.00	100.00	55.60
19	Credit Approval	82.13	51.69	84.06	85.51	83.57	52.66	84.06	68.60	83.09	85.99
20	Labor	94.12	82.35	88.24	82.35	94.12	82.35	76.47	94.12	82.35	88.24
21	Vote	97.69	96.92	90.77	96.15	90.00	96.15	96.15	97.69	92.31	96.15
22	Thoracic Surgery Data	84.40	83.69	84.40	83.69	82.27	82.98	76.60	80.85	75.18	83.69
23	Pen-Based Recognition	19.80	97.70	83.35	96.03	86.96	39.93	96.88	99.15	95.66	98.18
24	Letter Recognition	6.33	88.32	55.58	86.35	64.25	16.77	87.63	95.43	84.34	81.13
25	Statlog (Vehicle Silhouette)	39.58	75.35	63.54	67.01	44.44	51.39	66.67	76.74	66.67	73.26
26	Seismic-Bumps	93.29	92.77	93.29	93.29	82.06	25.03	93.29	0.00	92.00	91.87
27	Segment-Test	25.10	95.47	83.54	92.59	87.65	57.61	93.00	95.89	90.95	93.83
28	Segment-Challenge	30.00	96.44	83.78	96.44	81.11	65.33	94.67	98.22	92.67	90.67
29	Statlog (Credit Data)	71.50	96.90	70.90	70.00	87.30	97.20	70.00	100.00	100.00	97.00
30	Credit-G	75.00	74.33	73.33	73.67	75.33	71.67	76.67	77.33	67.00	75.67
31	Mushroom	96.26	100.00	93.51	100.00	95.89	98.52	100.00	100.00	100.00	100.00

Table 5.2(c): Output table by test (30%) (32-51 instances).

No	Datasets	AdaBoostM1	Bagging	IBK	J48	NaiveBayes	OneR	Part	RandomForest	RandomTree	SMO
32	Chronic Kidney Disease	98.33	99.17	85.00	100.00	94.17	92.50	96.67	100.00	96.67	97.50
33	Parkinsons	59.94	62.50	61.86	61.86	54.49	56.73	58.65	67.95	59.29	61.54
34	Higgs	70.10	89.03	55.36	86.13	59.28	71.04	71.57	100.00	60.70	63.60
35	Flags	39.66	36.21	56.90	44.83	46.55	5.17	51.72	60.34	51.72	62.07
36	Hypothyroid	92.67	99.38	92.31	99.21	95.23	95.76	98.94	99.12	98.06	93.29
37	Phishing Websites	92.82	96.02	90.56	95.60	93.09	88.75	95.84	97.23	95.84	93.61
38	Student Alcohol Consumption	29.86	37.56	15.84	34.84	38.46	41.63	37.10	30.32	18.55	21.27
39	Unbalanced	98.05	98.05	98.05	98.05	93.39	98.05	97.67	98.05	97.28	98.05
40	Alcohol Consumption	18.53	46.01	24.28	34.82	33.87	48.88	35.78	47.28	27.16	38.02
41	Ionosphere	82.86	84.76	73.33	80.00	80.95	73.33	82.86	90.48	82.86	85.71
42	Soybean	26.34	82.44	65.85	89.27	89.76	37.56	90.73	92.68	81.46	94.63
43	Annealing	83.63	97.40	90.98	99.78	86.64	83.63	100.00	100.00	100.00	99.00
44	Waveform	68.40	81.33	84.53	73.27	80.47	53.40	77.00	84.87	73.67	85.00
45	Kdd Cup 1999	99.90	99.97	99.87	99.97	99.90	99.87	100.00	100.00	100.00	100.00
46	Spambase	89.71	93.19	81.88	92.68	77.97	78.19	93.70	95.80	92.10	90.65
47	Optical Recognition	19.16	94.13	92.76	89.03	91.22	26.81	92.82	98.22	84.99	97.86
48	Bankruptcy	95.62	95.75	95.84	94.48	11.30	95.75	95.33	95.65	92.10	95.74
49	Supermarket	75.50	62.61	37.90	62.61	62.61	66.79	62.61	62.61	62.61	62.61
50	Multiple Features	17.33	84.17	89.83	78.50	93.67	22.00	80.67	96.83	65.33	97.83
51	Arrhythmia	57.35	76.47	55.88	71.32	70.59	58.09	63.97	71.32	54.41	78.68

5.1.2 Evaluation of Meta data

All the meta data for Bagging, Boosting, naive Bayes, K-Nearest neighbor, Decision tree, OneR, PART, Random Tree, Random Forest and Support Vector Machine are presented in Table 5.3 to 5.12.

Table 5.3: Meta data of Boosting.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	64.29	64.29	Average
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	50.00	64.29	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	70.83	62.50	Average
Iris	Small	Research	No	Low	Numeric	No	96.00	33.33	good
Balance Scale	Medium	Social	No	Low	Categorical	No	72.32	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	9.54	6.35	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	66.09	57.97	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	70.66	70.02	Average
Diabetes	Medium	Life	No	Low	Numeric	N/A	75.39	65.10	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	51.25	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	44.86	35.51	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	71.68	70.28	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	72.55	65.34	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	93.30	89.77	good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	78.09	64.84	Average
Wine Quality	Large	Business	No	Low	Real	No	93.42	42.59	good
Vowel Data	Medium	N/A	No	Low	Real	N/A	15.49	9.09	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	68.60	55.12	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	61.63	55.51	Average
Labor	Small	Social	No	Medium	Numeric	No	92.98	64.91	good
Vote	Small	Social	No	Medium	Nominal	N/A	95.63	61.38	good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	84.04	85.11	Average
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	20.38	10.41	Average
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	7.09	4.07	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	85.65	25.77	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	19.50	93.42	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	28.64	15.56	Average
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	30.40	15.73	Average
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	72.67	70.00	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	70.90	70.00	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	95.45	51.80	good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	99.00	62.50	good
Parkinsons	Large	Life	No	Medium	Real	N/A	51.03	50.00	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	66.40	52.29	Average
Flags	Small	N/A	No	Medium	Mixed	No	99.98	35.57	good
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	92.76	92.29	good
Phishing Websites	Big Data	Computer Security	No	High	Numeric	No	92.58	55.69	good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	39.95	16.02	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.60	good
Alcohol Consumption	Large	Social	No	High	Numeric	No	20.13	14.67	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	91.17	64.10	good
Soybean	Medium	Life	No	High	Nominal	No	27.97	13.47	Average
Annealing	Medium	Physical	No	High	Mixed	Yes	84.01	76.17	Average
Waveform	Large	Physical	Yes	High	Real	No	66.64	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.87	77.86	good
Spambase	Large	Computer	No	High	Mixed	Yes	90.07	60.60	good
Optical Recognition	Large	Computer	No	High	Numeric	No	19.75	10.18	Average
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.24	95.29	good
Supermarket	Large	Business	Yes	High	Numeric	N/A	75.38	63.71	Average
Multiple Features	Large	Computer	No	High	Mixed	No	27.74	10.00	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	55.53	54.20	Average

Table 5.4: Meta data of Bagging.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	57.14	64.29	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	57.14	64.29	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	66.67	62.50	Average
Iris	Small	Research	No	Low	Numeric	No	94.00	33.33	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	83.36	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.82	6.35	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	69.57	57.97	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	90.35	70.02	Good
Diabetes	Medium	Life	No	Low	Numeric	N/A	75.91	65.10	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	53.75	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	71.96	35.51	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	67.13	70.28	Bad
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	91.54	65.34	Good
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	97.50	89.77	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	86.96	64.84	Average
Wine Quality	Large	Business	No	Low	Real	No	65.17	42.59	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	98.59	9.09	Good
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	88.41	55.12	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	52.17	55.51	Average
Labor	Small	Social	No	Medium	Numeric	No	89.47	64.91	Average
Vote	Small	Social	No	Medium	Nominal	N/A	95.86	61.38	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	83.83	85.11	Average
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	97.91	10.41	Good
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	90.05	4.07	Good
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	72.10	25.77	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	92.45	93.42	Good
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	94.57	15.56	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	96.53	15.73	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	72.33	70.00	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	72.70	70.00	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	100.00	51.80	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	98.75	62.50	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	67.31	50.00	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	69.47	52.29	Average
Flags	Small	N/A	No	Medium	Mixed	No	35.57	35.57	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	99.47	92.29	Good
Phishing Websites	Big Data	Computer Secur	No	High	Numeric	No	96.20	55.69	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	45.45	16.02	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.60	Good
Alcohol Consumption	Large	Social	No	High	Numeric	No	46.88	14.67	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	92.59	64.10	Good
Soybean	Medium	Life	No	High	Nominal	No	84.63	13.47	Average
Annealing	Medium	Physical	No	High	Mixed	Yes	99.22	76.17	Good
Waveform	Large	Physical	Yes	High	Real	No	81.52	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.93	77.86	Good
Spambase	Large	Computer	No	High	Mixed	Yes	94.00	60.60	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	94.23	10.18	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.42	95.29	Good
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.71	Average
Multiple Features	Large	Computer	No	High	Mixed	No	84.00	10.00	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	72.12	54.20	Average

Table 5.5: Meta data of Naive Bayes.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	50.00	64.29	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	50.00	64.29	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	70.83	62.50	Average
Iris	Small	Research	No	Low	Numeric	No	96.00	33.33	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	90.40	46.08	Good
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.08	6.35	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	55.36	57.97	Bad
Car Evaluation	Large	N/A	No	Low	Categorical	No	88.42	70.02	Average
Diabetes	Medium	Life	No	Low	Numeric	N/A	75.78	65.10	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	58.75	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	47.66	35.51	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	72.73	70.28	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	69.62	65.34	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	82.16	89.77	Bad
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	72.66	64.84	Average
Wine Quality	Large	Business	No	Low	Real	No	55.03	42.59	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	58.59	9.09	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	45.17	55.12	Bad
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	84.93	55.51	Average
Labor	Small	Social	No	Medium	Numeric	No	94.74	64.91	Good
Vote	Small	Social	No	Medium	Nominal	N/A	90.12	61.38	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	78.51	85.11	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	85.75	10.41	Average
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	64.12	4.07	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	44.80	25.77	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	82.28	93.42	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	87.28	15.56	Average
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	81.33	15.73	Average
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	73.33	70.00	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	75.50	70.00	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	95.40	51.80	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	95.00	62.50	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	59.04	50.00	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	59.23	52.29	Average
Flags	Small	N/A	No	Medium	Mixed	No	55.15	35.57	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	95.39	92.29	Good
Phishing Websites	Big Data	Computer Secur	No	High	Numeric	No	92.98	55.69	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	42.68	16.02	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	89.60	98.60	Bad
Alcohol Consumption	Large	Social	No	High	Numeric	No	39.02	14.67	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	81.77	64.10	Average
Soybean	Medium	Life	No	High	Nominal	No	92.39	13.47	Good
Annealing	Medium	Physical	No	High	Mixed	Yes	87.73	76.17	Average
Waveform	Large	Physical	Yes	High	Real	No	80.00	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.82	77.86	Good
Spambase	Large	Computer	No	High	Mixed	Yes	79.29	60.60	Average
Optical Recognition	Large	Computer	No	High	Numeric	No	91.33	10.18	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	9.89	95.29	Bad
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.71	Average
Multiple Features	Large	Computer	No	High	Mixed	No	93.30	10.00	Good
Arrhythmia	Small	Life	No	High	Mixed	Yes	62.39	54.20	Average

Table 5.6: Meta data of K-Nearest neighbors.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	57.14	64.29	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	92.86	64.29	Good
Contact-Lenses	Small	Life	No	Low	Nominal	No	75.00	62.50	Average
Iris	Small	Research	No	Low	Numeric	No	96.67	33.33	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	89.44	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.31	6.35	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	64.35	57.97	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	71.24	70.02	Average
Diabetes	Medium	Life	No	Low	Numeric	N/A	75.00	65.10	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	52.50	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	62.15	35.51	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	73.43	70.28	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	75.47	65.34	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	89.46	89.77	Bad
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	78.38	64.84	Average
Wine Quality	Large	Business	No	Low	Real	No	61.04	42.59	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	33.00	9.09	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	65.23	55.12	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	85.65	55.51	Average
Labor	Small	Social	No	Medium	Numeric	No	91.23	64.91	Good
Vote	Small	Social	No	Medium	Nominal	N/A	91.26	61.38	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	85.11	85.11	Average
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	84.89	10.41	Average
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	57.91	4.07	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	64.66	25.77	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	93.42	93.42	Good
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	85.80	15.56	Average
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	85.47	15.73	Average
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	72.67	70.00	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	72.30	70.00	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	89.66	51.80	Average
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	85.50	62.50	Average
Parkinsons	Large	Life	No	Medium	Real	N/A	65.29	50.00	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	54.67	52.29	Average
Flags	Small	N/A	No	Medium	Mixed	No	66.49	35.57	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	92.29	92.29	Average
Phishing Websites	Big Data	Computer Security	No	High	Numeric	No	90.76	55.69	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	14.18	16.02	Bad
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.60	Average
Alcohol Consumption	Large	Social	No	High	Numeric	No	28.86	14.67	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	84.90	64.10	Average
Soybean	Medium	Life	No	High	Nominal	No	68.67	13.47	Average
Annealing	Medium	Physical	No	High	Mixed	Yes	89.59	76.17	Average
Waveform	Large	Physical	Yes	High	Real	No	84.96	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.76	77.86	Good
Spambase	Large	Computer	No	High	Mixed	Yes	83.26	60.60	Average
Optical Recognition	Large	Computer	No	High	Numeric	No	93.86	10.18	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.29	95.29	Average
Supermarket	Large	Business	Yes	High	Numeric	N/A	37.07	63.71	Bad
Multiple Features	Large	Computer	No	High	Mixed	No	89.95	10.00	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	55.53	54.20	Average

Table 5.7: Meta data of Decision tree.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	42.86	64.29	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	57.14	64.29	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	83.33	62.50	Average
Iris	Small	Research	No	Low	Numeric	No	94.00	33.33	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	76.64	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.62	6.35	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	68.70	57.97	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	90.93	70.02	Good
Diabetes	Medium	Life	No	Low	Numeric	N/A	72.01	65.10	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	58.75	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	64.49	35.51	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	73.43	70.28	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	84.55	65.34	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	97.44	89.77	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	85.08	64.84	Average
Wine Quality	Large	Business	No	Low	Real	No	61.48	42.59	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	77.78	9.09	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	82.31	55.12	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	85.22	55.51	Average
Labor	Small	Social	No	Medium	Numeric	No	82.46	64.91	Average
Vote	Small	Social	No	Medium	Nominal	N/A	96.32	61.38	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	84.47	85.11	Average
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	96.56	10.41	Good
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	87.98	4.07	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	72.46	25.77	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	93.42	93.42	Good
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	93.33	15.56	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	95.80	15.73	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	73.67	70.00	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	70.90	70.00	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	100.00	51.80	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	99.00	62.50	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	62.79	50.00	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	65.17	52.29	Average
Flags	Small	N/A	No	Medium	Mixed	No	59.28	35.57	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	99.60	92.29	Good
Phishing Websites	Big Data	Computer Security	No	High	Numeric	No	95.88	55.69	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	39.45	16.02	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.60	Good
Alcohol Consumption	Large	Social	No	High	Numeric	No	39.41	14.67	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	91.17	64.10	Good
Soybean	Medium	Life	No	High	Nominal	No	91.65	13.47	Good
Annealing	Medium	Physical	No	High	Mixed	Yes	98.88	76.17	Good
Waveform	Large	Physical	Yes	High	Real	No	75.08	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.89	77.86	Good
Spambase	Large	Computer	No	High	Mixed	Yes	92.98	60.60	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	90.69	10.18	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.13	95.29	Good
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.71	Average
Multiple Features	Large	Computer	No	High	Mixed	No	78.65	10.00	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	64.38	54.20	Average

Table 5.8: Meta data of OneR.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	50.00	64.2857	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	42.86	64.2857	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	70.83	62.5	Average
Iris	Small	Research	No	Low	Numeric	No	92.00	33.3333	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	56.32	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.62	6.3476	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	55.07	57.971	Bad
Car Evaluation	Large	N/A	No	Low	Categorical	No	70.66	70.0231	Average
Diabetes	Medium	Life	No	Low	Numeric	N/A	71.48	65.1042	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	50.00	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	56.54	35.514	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	65.39	70.2797	Bad
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	69.94	65.3445	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	94.34	89.768	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	70.80	64.8352	Average
Wine Quality	Large	Business	No	Low	Real	No	54.66	42.5891	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	29.97	9.0909	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	63.53	55.1202	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	85.51	55.5072	Average
Labor	Small	Social	No	Medium	Numeric	No	73.68	64.9123	Average
Vote	Small	Social	No	Medium	Nominal	N/A	95.63	61.3793	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	83.40	85.1064	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	38.95	10.4076	Average
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	17.24	4.065	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	51.89	25.7683	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	28.10	93.4211	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	60.99	15.5556	Average
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	64.53	15.7333	Average
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	72.67	70	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	66.40	70	Bad
Mushroom	Large	Life	No	Medium	Categorical	Yes	98.69	51.7971	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	92.00	62.5	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	56.44	50	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	58.35	52.285	Average
Flags	Small	N/A	No	Medium	Mixed	No	4.64	35.567	Bad
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	96.00	92.2853	Good
Phishing Websites	Big Data	Computer Secur	No	High	Numeric	No	88.89	55.6943	Average
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	46.23	16.0247	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.5981	Good
Alcohol Consumption	Large	Social	No	High	Numeric	No	49.86	14.6692	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	80.63	64.1026	Average
Soybean	Medium	Life	No	High	Nominal	No	39.97	13.47	Average
Annealing	Medium	Physical	No	High	Mixed	Yes	84.01	76.1693	Average
Waveform	Large	Physical	Yes	High	Real	No	53.72	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.90	77.8622	Good
Spambase	Large	Computer	No	High	Mixed	Yes	78.33	60.5955	Average
Optical Recognition	Large	Computer	No	High	Numeric	No	27.03	10.1779	Average
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.19	95.2871	Bad
Supermarket	Large	Business	Yes	High	Numeric	N/A	66.74	63.713	Average
Multiple Features	Large	Computer	No	High	Mixed	No	23.10	10	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	60.40	54.2035	Average

Table 5.9: Meta data of PART.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	50.00	64.2857	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	35.71	64.2857	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	83.33	62.5	Average
Iris	Small	Research	No	Low	Numeric	No	94.00	33.3333	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	83.52	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.62	6.3476	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	63.77	57.971	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	94.59	70.0231	Good
Diabetes	Medium	Life	No	Low	Numeric	N/A	74.74	65.1042	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	58.75	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	69.63	35.514	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	66.43	70.2797	Bad
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	94.26	65.3445	Good
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	97.02	89.768	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	85.20	64.8352	Average
Wine Quality	Large	Business	No	Low	Real	No	61.35	42.5891	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	68.01	9.0909	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	63.53	55.1202	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	83.33	55.5072	Average
Labor	Small	Social	No	Medium	Numeric	No	80.70	64.9123	Average
Vote	Small	Social	No	Medium	Nominal	N/A	96.55	61.3793	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	79.15	85.1064	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	96.83	10.4076	Good
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	88.76	4.065	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	71.51	25.7683	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	93.42	93.4211	Good
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	92.35	15.5556	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	95.60	15.7333	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	73.67	70	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	69.40	70	Bad
Mushroom	Large	Life	No	Medium	Categorical	Yes	100.00	51.7971	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	98.50	62.5	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	59.23	50	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	65.17	52.285	Average
Flags	Small	N/A	No	Medium	Mixed	No	57.73	35.567	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	99.52	92.2853	Good
Phishing Websites	Big Data	Computer Secur	No	High	Numeric	No	96.76	55.6943	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	34.82	16.0247	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	97.78	98.5981	Bad
Alcohol Consumption	Large	Social	No	High	Numeric	No	40.17	14.6692	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	90.60	64.1026	Good
Soybean	Medium	Life	No	High	Nominal	No	91.22	13.47	Good
Annealing	Medium	Physical	No	High	Mixed	Yes	98.51	76.1693	Good
Waveform	Large	Physical	Yes	High	Real	No	77.42	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.94	77.8622	Good
Spambase	Large	Computer	No	High	Mixed	Yes	93.59	60.5955	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	92.54	10.1779	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.22	95.2871	Good
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.713	Average
Multiple Features	Large	Computer	No	High	Mixed	No	82.00	10	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	63.94	54.2035	Average

Table 5.10: Meta data of Random Forest.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	71.43	64.2857	Average
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	42.86	64.2857	Average
Contact-Lenses	Small	Life	No	Low	Nominal	No	70.83	62.5	Average
Iris	Small	Research	No	Low	Numeric	No	94.00	33.3333	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	81.44	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	0.00	6.3476	Bad
Liver Disorders	Small	Life	No	Low	Mixed	No	73.04	57.971	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	93.82	70.0231	Good
Diabetes	Medium	Life	No	Low	Numeric	N/A	75.91	65.1042	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	61.25	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	80.37	35.514	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	68.53	70.2797	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	96.35	65.3445	Good
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	97.50	89.768	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	88.01	64.8352	Average
Wine Quality	Large	Business	No	Low	Real	No	70.11	42.5891	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	93.60	9.0909	Good
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	92.52	55.1202	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	75.51	55.5072	Average
Labor	Small	Social	No	Medium	Numeric	No	91.23	64.9123	Good
Vote	Small	Social	No	Medium	Nominal	N/A	96.32	61.3793	Average
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	83.62	85.1064	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	99.13	10.4076	Good
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	96.51	4.065	Good
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	76.00	25.7683	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	0.00	93.4211	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	95.80	15.5556	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	97.93	15.7333	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	75.67	70	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	75.70	70	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	100.00	51.7971	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	100.00	62.5	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	70.10	50	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	71.57	52.285	Average
Flags	Small	N/A	No	Medium	Mixed	No	61.34	35.567	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	99.39	92.2853	Good
Phishing Websites	Big Data	Computer Security	No	High	Numeric	No	97.26	55.6943	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	37.29	16.0247	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.5981	Good
Alcohol Consumption	Large	Social	No	High	Numeric	No	45.06	14.6692	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	93.73	64.1026	Good
Soybean	Medium	Life	No	High	Nominal	No	92.97	13.47	Good
Annealing	Medium	Physical	No	High	Mixed	Yes	98.88	76.1693	Good
Waveform	Large	Physical	Yes	High	Real	No	85.06	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.97	77.8622	Good
Spambase	Large	Computer	No	High	Mixed	Yes	95.50	60.5955	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	98.15	10.1779	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.02	95.2871	Good
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.713	Average
Multiple Features	Large	Computer	No	High	Mixed	No	96.50	10	Good
Arrhythmia	Small	Life	No	High	Mixed	Yes	67.26	54.2035	Average

Table 5.11: Meta data of Random Tree.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	57.14	64.2857	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	35.71	64.2857	Bad
Contact-Lenses	Small	Life	No	Low	Nominal	No	75.00	62.5	Average
Iris	Small	Research	No	Low	Numeric	No	94.67	33.3333	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	77.28	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.49	6.3476	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	67.82	57.971	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	85.52	70.0231	Average
Diabetes	Medium	Life	No	Low	Numeric	N/A	67.84	65.1042	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	58.75	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	69.16	35.514	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	62.94	70.2797	Bad
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	79.23	65.3445	Average
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	96.41	89.768	Good
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	81.60	64.8352	Average
Wine Quality	Large	Business	No	Low	Real	No	62.41	42.5891	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	81.14	9.0909	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	81.98	55.1202	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	62.61	55.5072	Average
Labor	Small	Social	No	Medium	Numeric	No	84.21	64.9123	Average
Vote	Small	Social	No	Medium	Nominal	N/A	94.48	61.3793	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	74.89	85.1064	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	95.72	10.4076	Average
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	86.16	4.065	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	70.92	25.7683	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	91.95	93.4211	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	91.36	15.5556	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	95.07	15.7333	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	70.33	70	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	66.10	70	Bad
Mushroom	Large	Life	No	Medium	Categorical	Yes	99.96	51.7971	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	95.50	62.5	Good
Parkinsons	Large	Life	No	Medium	Real	N/A	63.08	50	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	60.70	52.285	Average
Flags	Small	N/A	No	Medium	Mixed	No	42.78	35.567	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	97.64	92.2853	Good
Phishing Websites	Big Data	Computer Secur	No	High	Numeric	No	96.37	55.6943	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	30.05	16.0247	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	97.55	98.5981	Bad
Alcohol Consumption	Large	Social	No	High	Numeric	No	30.97	14.6692	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	90.60	64.1026	Good
Soybean	Medium	Life	No	High	Nominal	No	83.90	13.47	Average
Annealing	Medium	Physical	No	High	Mixed	Yes	97.40	76.1693	Good
Waveform	Large	Physical	Yes	High	Real	No	73.08	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.95	77.8622	Good
Spambase	Large	Computer	No	High	Mixed	Yes	90.94	60.5955	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	85.94	10.1779	Average
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	92.05	95.2871	BAD
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.713	Average
Multiple Features	Large	Computer	No	High	Mixed	No	65.35	10	Average
Arrhythmia	Small	Life	No	High	Mixed	Yes	51.11	54.2035	Bad

Table 5.12: Meta data of Support Vector Machine.

Dataset Name	Size of Dataset	Area	Streaming Data	Dimension	Attribute Type	Missing Data	Accuracy	ZeroR	Class
Weather Nominal	Small	Physical	Yes	Low	Nominal	No	57.14	64.286	Bad
Weather Numeric	Small	Physical	Yes	Low	Numeric	No	64.29	64.286	Average
Contact-Lenses	Small	Life	No	Low	Nominal	No	66.67	62.5	Average
Iris	Small	Research	No	Low	Numeric	No	96.00	33.333	Good
Balance Scale	Medium	Social	No	Low	Categorical	No	87.68	46.08	Average
Tamilnadu Electricity Board	Big Data	Life	Yes	Low	Mixed	N/A	65.31	6.3476	Average
Liver Disorders	Small	Life	No	Low	Mixed	No	58.26	57.971	Average
Car Evaluation	Large	N/A	No	Low	Categorical	No	93.63	70.023	Good
Diabetes	Medium	Life	No	Low	Numeric	N/A	76.95	65.104	Average
Gps Trajectories	Small	Computer	Yes	Low	Real	Yes	52.50	41.25	Average
Glass	Small	Physical	No	Low	Numeric	No	56.54	35.514	Average
Breast-Cancer	Small	Life	No	Low	Nominal	No	70.28	70.28	Average
Tic Tac Toe Endgame	Medium	Game	No	Low	Categorical	No	98.33	65.345	Good
Page Blocks Classification	Large	Computer	No	Low	Mixed	No	92.08	89.768	Average
Magic Gamma Telescope	Big Data	Physical	No	Low	Real	N/A	79.14	64.835	Average
Wine Quality	Large	Business	No	Low	Real	No	58.35	42.589	Average
Vowel Data	Medium	N/A	No	Low	Real	N/A	61.28	9.0909	Average
Eeg Eye State	Big Data	Life	No	Low	Mixed	N/A	55.61	55.12	Average
Credit Approval	Medium	Business	No	Medium	Mixed	Yes	86.67	55.507	Average
Labor	Small	Social	No	Medium	Numeric	No	94.74	64.912	Good
Vote	Small	Social	No	Medium	Nominal	N/A	96.09	61.379	Good
Thoracic Surgery Data	Small	Life	Yes	Medium	Mixed	No	84.89	85.106	Bad
Pen-Based Recognition	Big Data	Computer	No	Medium	Numeric	No	97.96	10.408	Good
Letter Recognition	Big Data	Computer	No	Medium	Numeric	No	82.34	4.065	Average
Statlog (Vehicle Silhouettes)	Medium	N/A	No	Medium	Numeric	No	74.35	25.768	Average
Seismic-Bumps	Medium	N/A	No	Medium	Real	N/A	91.83	93.421	Bad
Segment-Test	Medium	N/A	N/A	Medium	Numeric	N/A	91.98	15.556	Good
Segment-Challenge	Large	N/A	N/A	Medium	Numeric	N/A	91.67	15.733	Good
Statlog (German Credit Data)	Medium	Financial	No	Medium	Mixed	No	75.00	70	Average
Credit-G	Medium	Financial	No	Medium	Nominal	N/A	75.10	70	Average
Mushroom	Large	Life	No	Medium	Categorical	Yes	100.00	51.797	Good
Chronic Kidney Disease	Small	N/A	Yes	Medium	Real	Yes	97.75	62.5	Average
Parkinsons	Large	Life	No	Medium	Real	N/A	63.08	50	Average
Higgs	Big Data	Physical	No	Medium	Numeric	N/A	63.60	52.285	Average
Flags	Small	N/A	No	Medium	Mixed	No	60.82	35.567	Average
Hypothyroid	Large	Life	No	Medium	Numeric	N/A	93.58	92.285	Good
Phishing Websites	Big Data	Computer Security	No	High	Numeric	No	93.80	55.694	Good
Student Alcohol Consumption	Medium	Social	No	High	Numeric	N/A	22.65	16.025	Average
Unbalanced	Medium	N/A	No	High	Numeric	N/A	98.60	98.598	Average
Alcohol Consumption	Large	Social	No	High	Numeric	No	40.27	14.669	Average
Ionosphere	Small	Physical	Yes	High	Numeric	No	87.46	64.103	Average
Soybean	Medium	Life	No	High	Nominal	No	92.83	13.47	Good
Annealing	Medium	Physical	No	High	Mixed	Yes	96.65	76.169	Good
Waveform	Large	Physical	Yes	High	Real	No	86.68	33.84	Average
Kdd Cup 1999	Big Data	Computer	No	High	Mixed	No	99.98	77.862	Good
Spambase	Large	Computer	No	High	Mixed	Yes	90.42	60.596	Good
Optical Recognition	Large	Computer	No	High	Numeric	No	98.33	10.178	Good
Polish Companies Bankruptcy	Big Data	Business	Yes	High	Real	Yes	95.28	95.287	Bad
Supermarket	Large	Business	Yes	High	Numeric	N/A	63.71	63.713	Average
Multiple Features	Large	Computer	No	High	Mixed	No	97.55	10	Good
Arrhythmia	Small	Life	No	High	Mixed	Yes	70.13	54.204	Average

5.2 Extracting Meta-Knowledge

In chapter 4, we described that a meta data can be generated by extracting previous knowledge. A meta-knowledge is a decision tree or model of a decision in which each internal nodes represent a test feature of meta data, each branch represent an outcome of the test and each leaf node represent a class label of meta data [3].

For the test cases, we present meta-knowledge sub-trees related to the Life area and Social area in Figure 5.1 to 5.20.

1. Meta-knowledge (Life area) of K-nearest neighbor

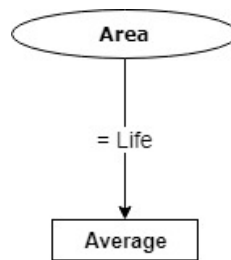


Figure 5.1: Meta-knowledge (sub-tree) of K-nearest neighbor.

2. Meta-knowledge (Life area) of Boosting

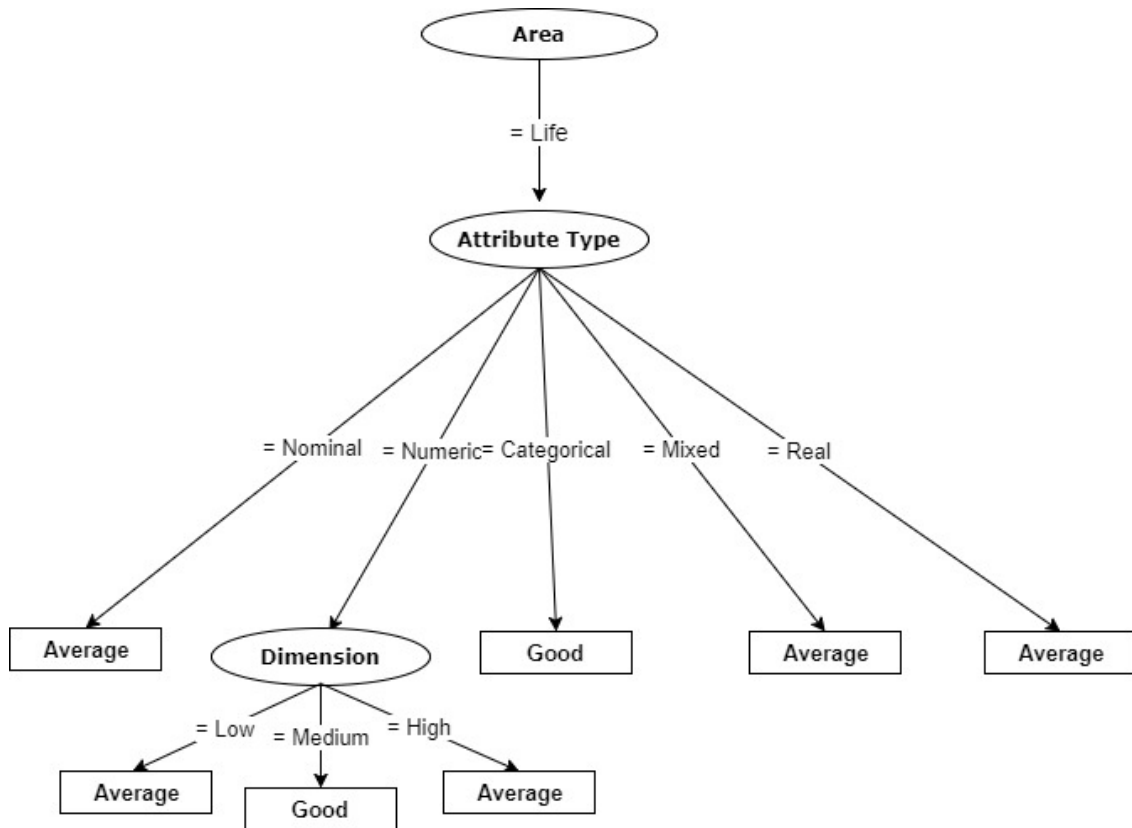


Figure 5.2: Meta-knowledge (sub-tree) of Boosting.

3. Meta-knowledge (Life area) of Bagging

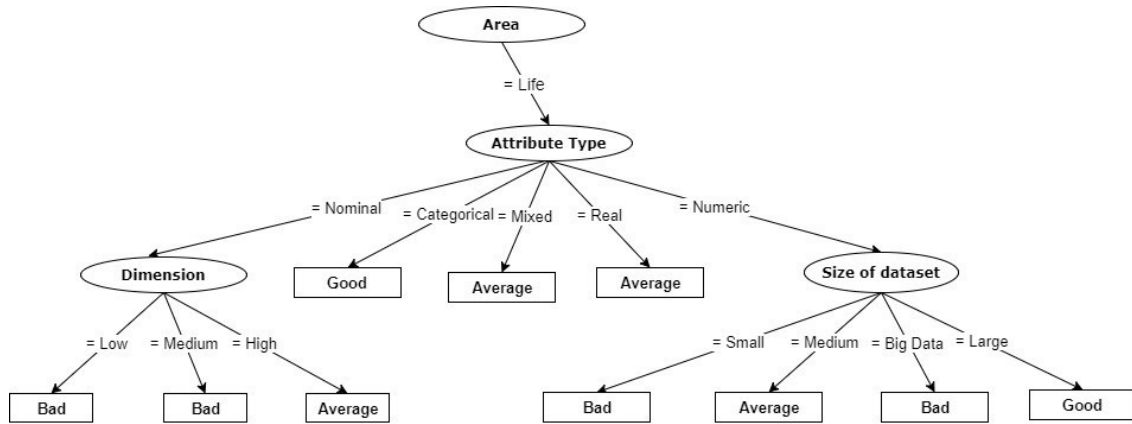


Figure 5.3: Meta-knowledge (sub-tree) of Bagging.

4. Meta-knowledge (Life area) of Decision Tree

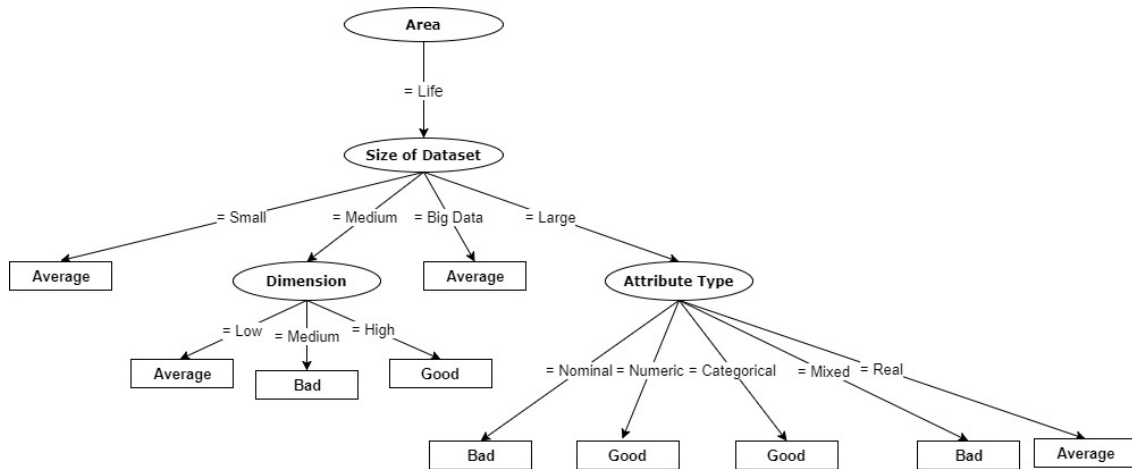


Figure 5.4: Meta-knowledge (sub-tree) of Decision Tree.

5. Meta-knowledge (Life area) of Naive Bayes

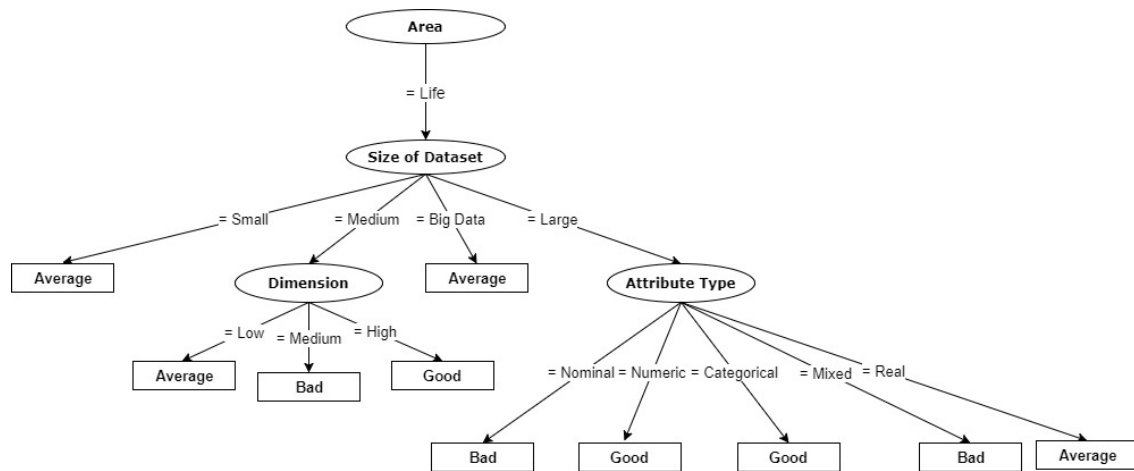


Figure 5.5: Meta-knowledge (sub-tree) of Naive Bayes.

6. Meta-knowledge (Life area) of OneR

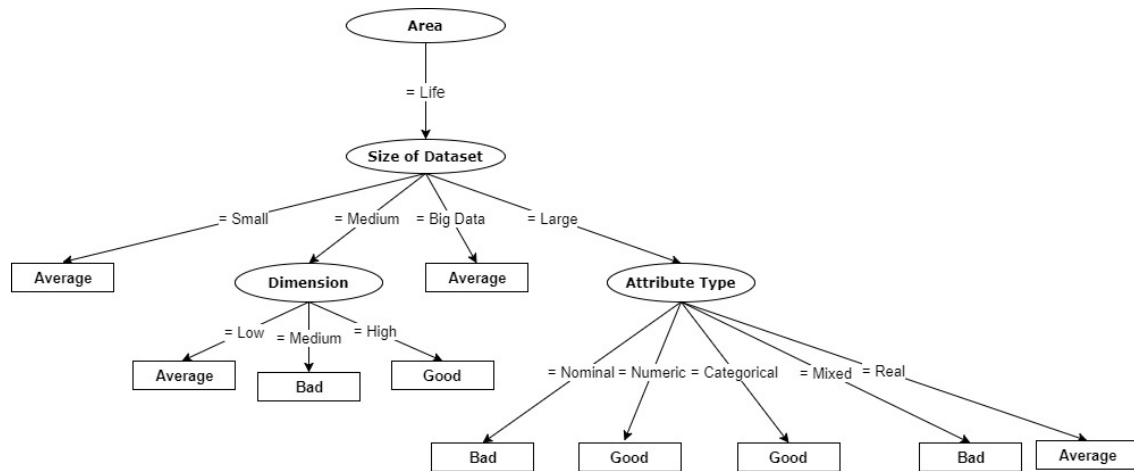


Figure 5.6: Meta-knowledge (sub-tree) of OneR.

7. Meta-knowledge (Life area) of Part

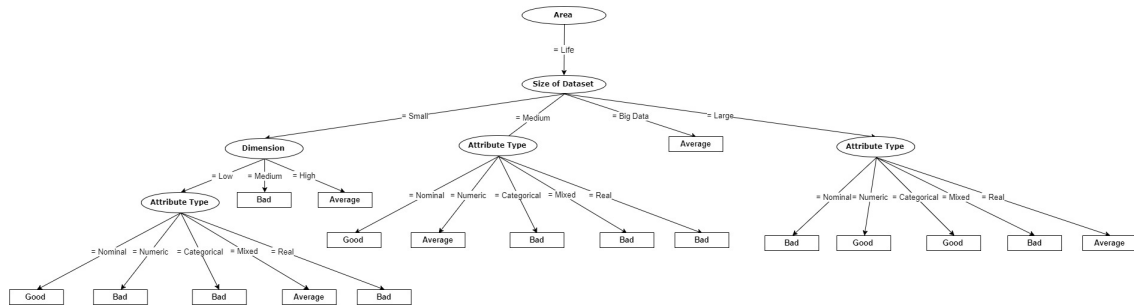


Figure 5.7: Meta-knowledge (sub-tree) of Part.

8. Meta-knowledge (Life area) of Random Forest

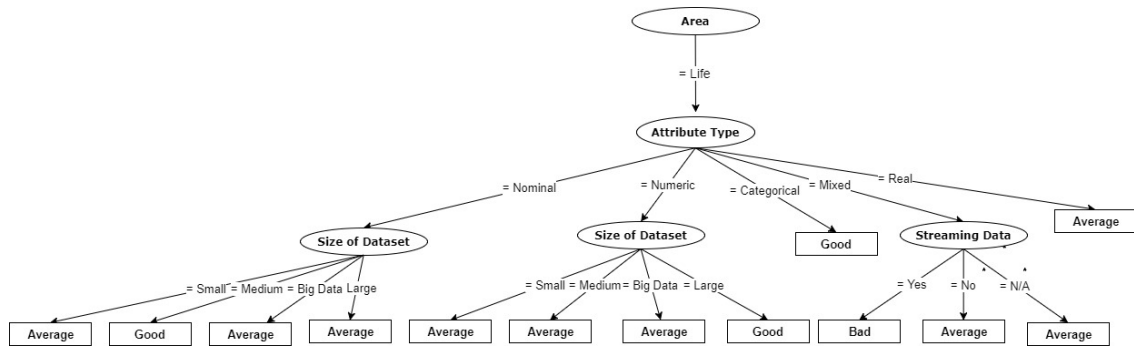


Figure 5.8: Meta-knowledge (sub-tree) of Random Forest.

9. Meta-knowledge (Life area) of Random Tree

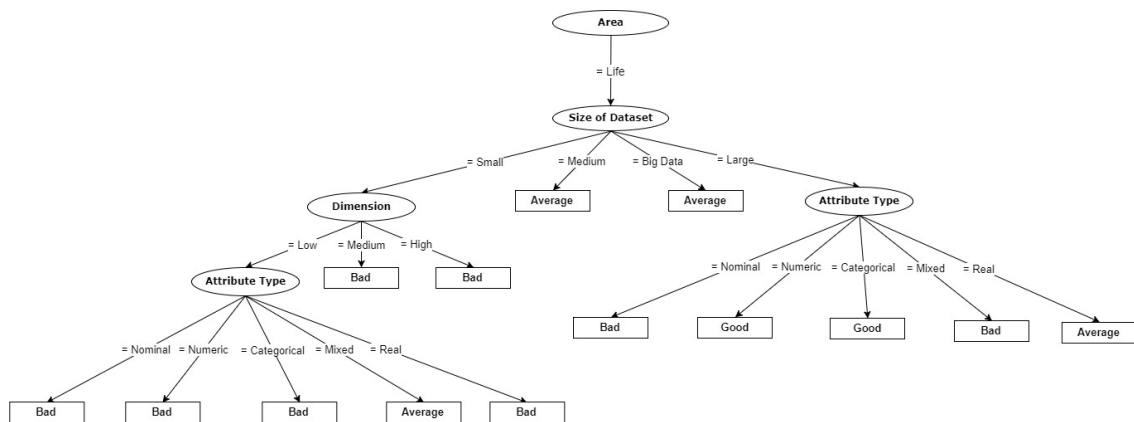


Figure 5.9: Meta-knowledge (sub-tree) of Random Tree.

10. Meta-knowledge (Life area) of Support Vector Machine

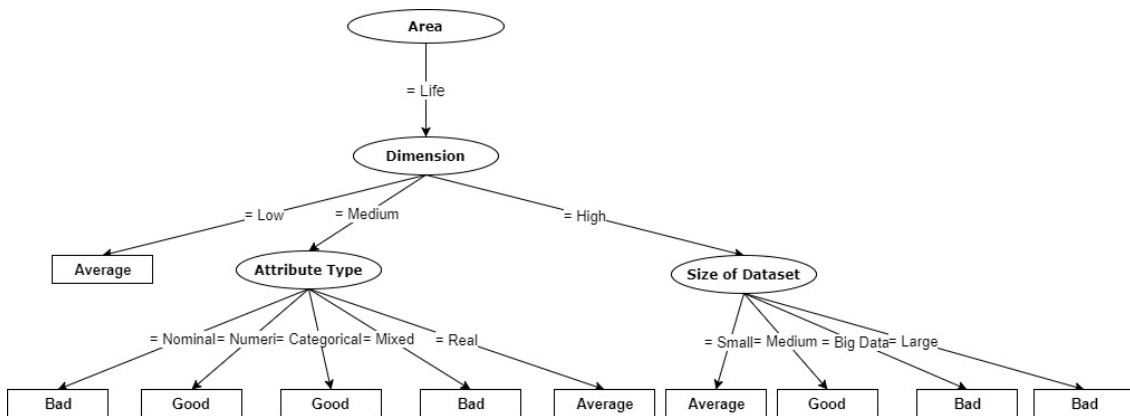


Figure 5.10: Meta-knowledge (sub-tree) of Support Vector Machine.

11. Meta-knowledge (Social area) of K-nearest neighbor

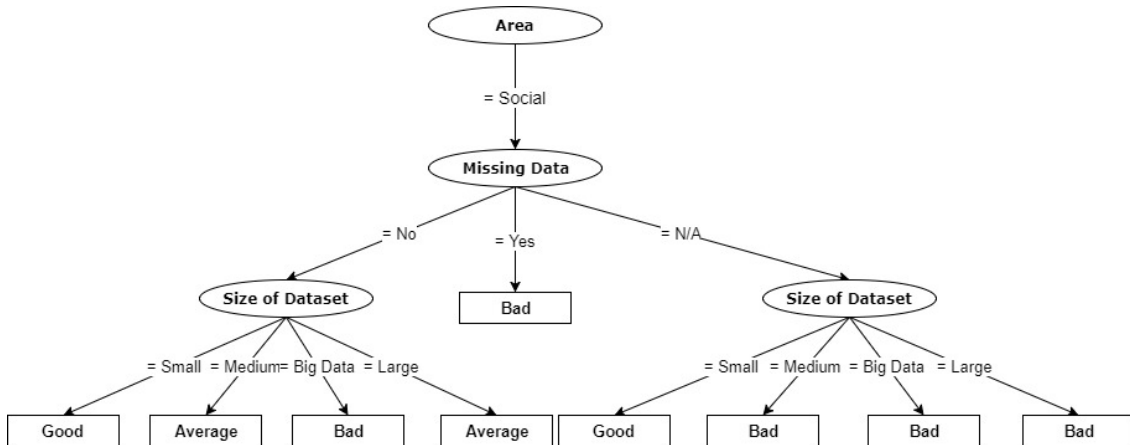


Figure 5.11: Meta-knowledge (sub-tree) of K-nearest neighbor.

12. Meta-knowledge (Social area) of Boosting

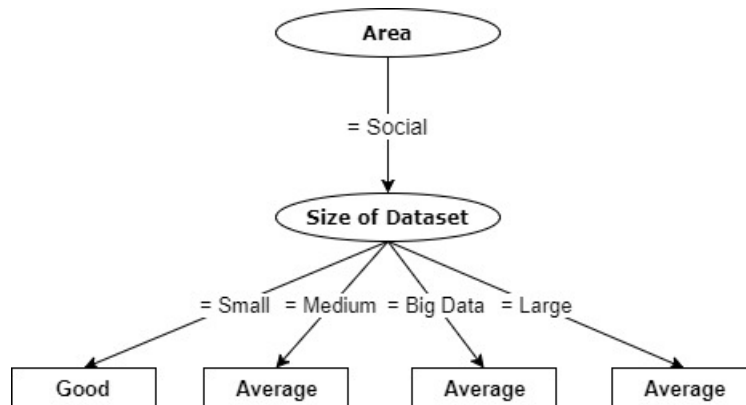


Figure 5.12: Meta-knowledge (sub-tree) of Boosting.

13. Meta-knowledge (Social area) of Baggingr

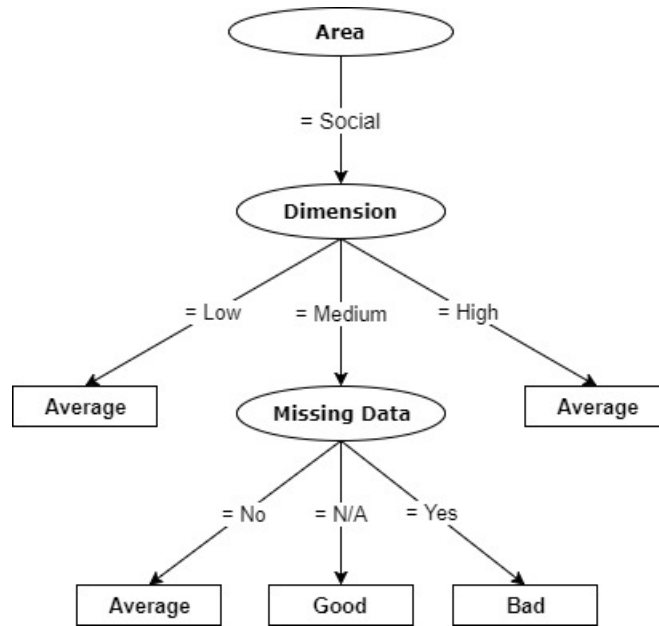


Figure 5.13: Meta-knowledge (sub-tree) of Bagging.

14. Meta-knowledge (Social area) of Decision Tree

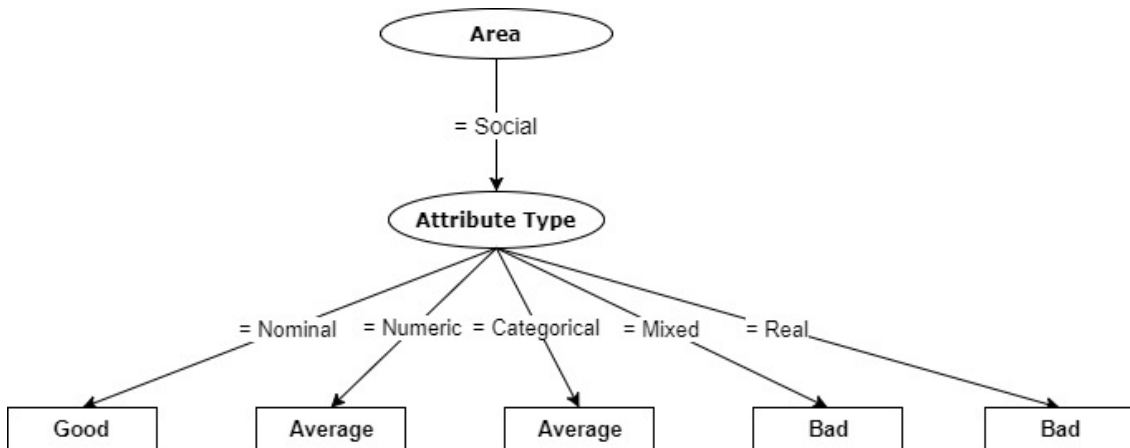


Figure 5.14: Meta-knowledge (sub-tree) of Decision Tree.

15. Meta-knowledge (Social area) of Naive Bayes

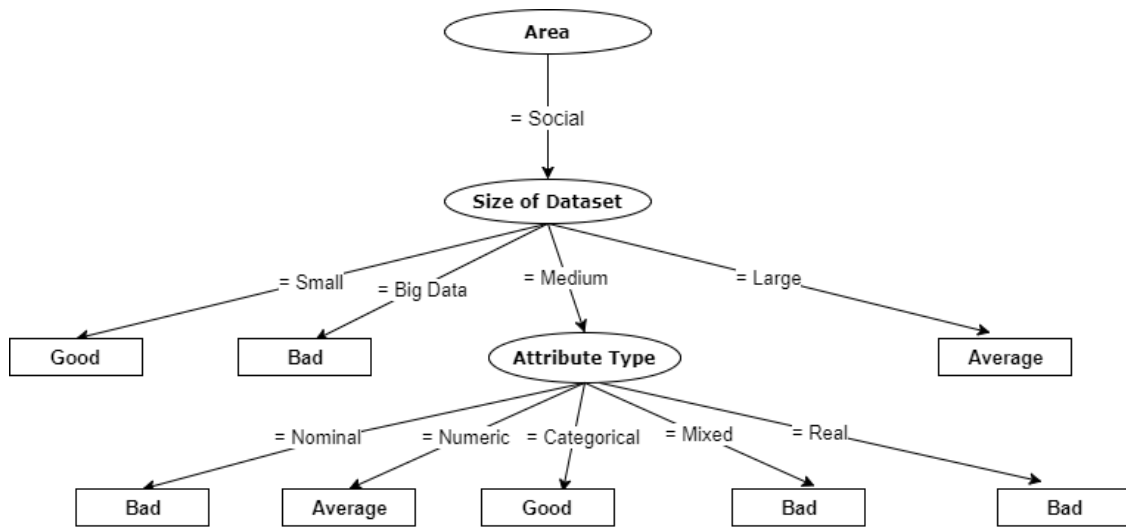


Figure 5.15: Meta-knowledge (sub-tree) of Naive Bayes.

16. Meta-knowledge (Social area) of OneR

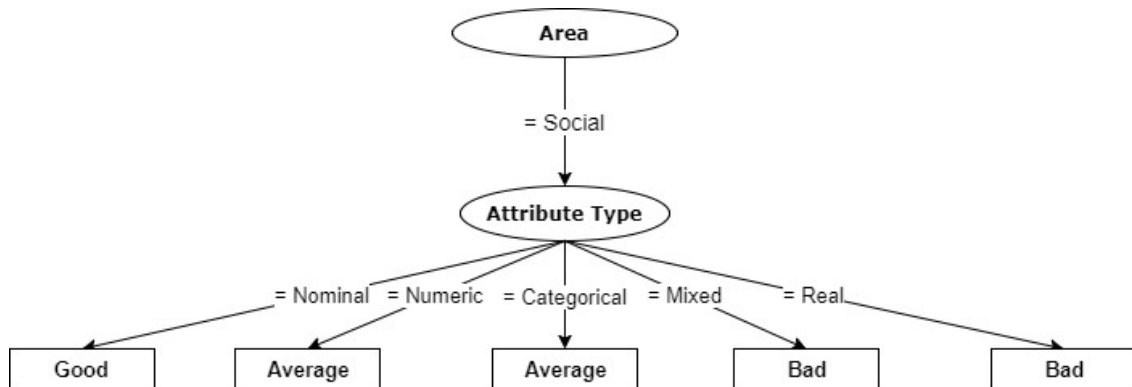


Figure 5.16: Meta-knowledge (sub-tree) of OneR.

17. Meta-knowledge (Social area) of Part

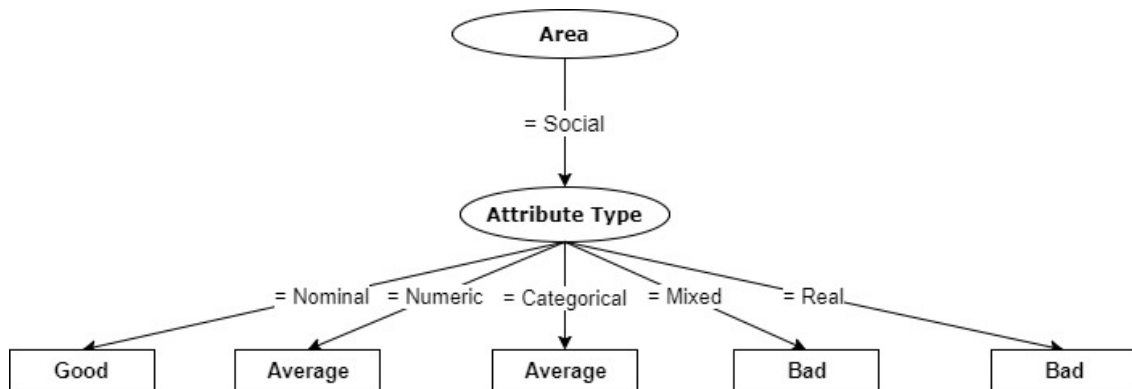


Figure 5.17: Meta-knowledge (sub-tree) of Part.

18. Meta-knowledge (Social area) of Random Forest

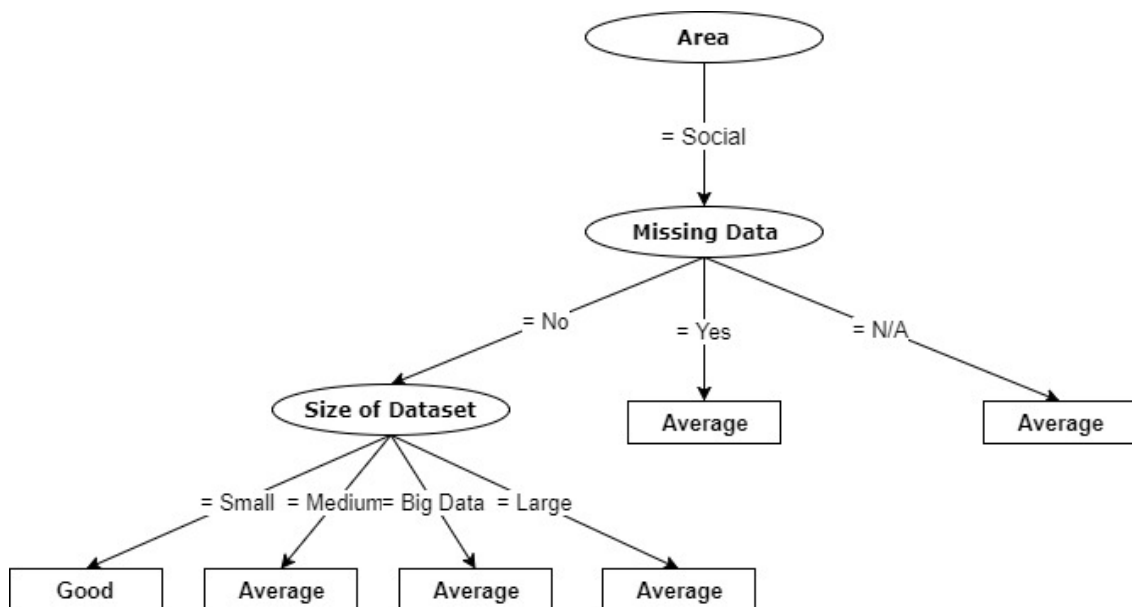


Figure 5.18: Meta-knowledge (sub-tree) of Random Forest.

19. Meta-knowledge (Social area) of Random Tree

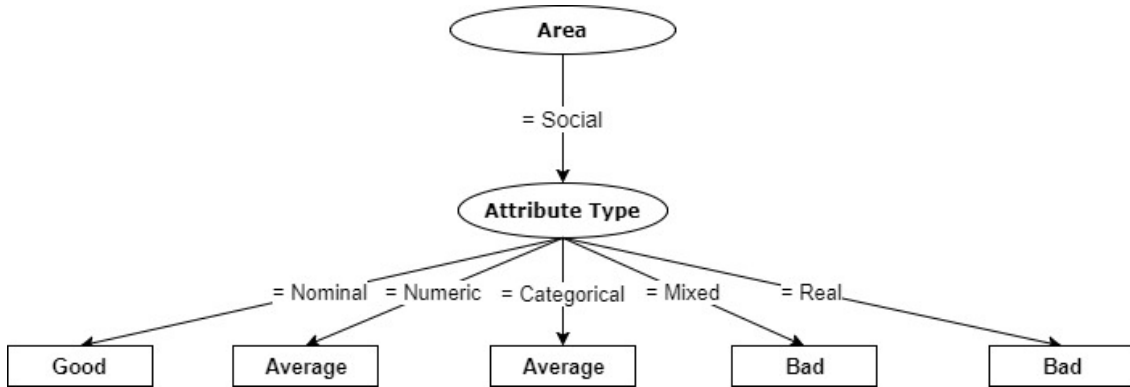


Figure 5.19: Meta-knowledge (sub-tree) of Random Tree.

20. Meta-knowledge (Social area) of Support Vector Machine

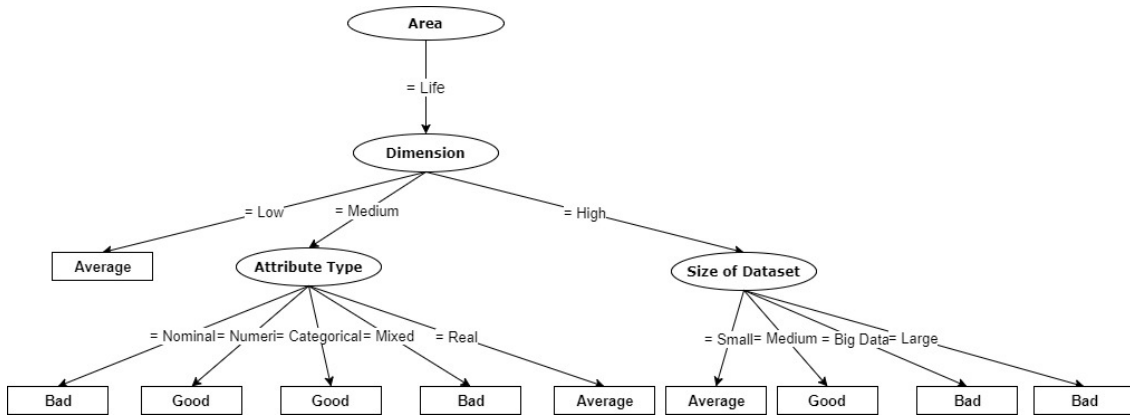


Figure 5.20: Meta-knowledge (sub-tree) of Support Vector Machine.

5.3 Performance Analysis

To analyze the performance of the extracting meta knowledge, we generated comparison tables, where the table consists of meta-knowledge decision and evaluation decision generated by Weka. First of all, we evaluated the extracting meta knowledge using the datasets and generated the decision for the comparison table. Hence, to get evaluation decision for the comparison table, we evaluated the datasets by cross-validation. We illustrated comparison tables in Table 5.14 to 5.18. The output of the comparison table described in the test case 1 to 5.

To compare the performance we are going to use datasets from UCI machine learning repository. Table 5.13 describes the datasets descriptions.

TABLE 5.13: Test Dataset Description.

Dataset Name	Area	Size of Dataset	Dimension	Attribute Type	Streaming Data	Missing Data
Anuran	Life	Large	Medium	Numeric	Yes	N/A
Molecular	Life	Large	Low	Nominal	No	No
Yeast	Life	Large	Low	Nominal	No	No
Ballons	Social	Small	Low	Nominal	No	No
Hayes-Roth	Social	Small	Low	Numeric	No	No

Test 1:

For the test 1, we selected anuran dataset which is from Life area, has 7195 instances, 26 attributes, Attribute type is Numeric also it has no missing value and it has the streaming data source. After converting the information into meta information, Size of the dataset is “Large”, Dimension is “Medium”, Attribute type is “Numeric”, streaming data is “Yes” and the missing data feature is “N/A”.

Algorithm Name	Accuracy (Weka)	Standard Accuracy (ZeroR)	Decision (Weka)	Decision (Meta Knowledge)
Naive Bayes	99.07%	48.34%	Good	Good
Decision Tree	100%	48.34%	Good	Good
Random Tree	99.42%	48.34%	Good	Good
Random Forest	100%	48.34%	Good	Good
Bagging	100%	48.34%	Good	Good
Boosting (Adaboost)	63.92%	48.34%	Average	Good
OneR	99.99%	48.34%	Good	Good
PART	100%	48.34%	Good	Good
Support Vector Machine(SVM)	99.89%	48.34%	Good	Good
K-Nearest Neighbor(KNN)	99.82%	48.34%	Good	Average

Table 5.14: Performance comparison of Anuran dataset.

According to the Table 5.14, the decision of our meta-knowledge for naive bayes classifier is “Good” that means this naive bayes classifier can be able to classify anuran test dataset with more than 90% accuracy where standard accuracy is 48.34%. In weka, the decision is also “Good” because naive bayes correctly classified 99.07% using ten-fold cross-validation. So in this domain, the prediction of our meta-knowledge is 100% accurate. We also measured that the prediction of our meta-knowledge is almost 98% accurate for this test dataset.

Test 2:

For the test 2, we selected molecular dataset which is from Life area, has 3190 instances, 3 attributes, Attribute type is Nominal also it has no missing value and streaming data. After converting the information into meta information, size of the dataset is “Large”, Dimension is “Low”, Attribute type is “Nominal”, streaming data is “No” and the missing data feature is “No”.

According to the Table 5.15, the decision of meta-knowledge for support vector classifier is “Average” that means this support vector classifier can be able to classify this test dataset with more than standard accuracy but it will have less than 90% classifier accuracy. In weka, the

Algorithm Name	Accuracy (Weka)	Standard Accuracy (ZeroR)	Decision (Weka)	Decision (Meta Knowledge)
Naive Bayes	100%	60%	Good	Good
Decision Tree	100%	60%	Good	Good
Random Tree	100%	60%	Good	Good
Random Forest	100%	60%	Good	Good
Bagging	90%	60%	Good	Average
Boosting (Adaboost)	100%	60%	Good	Good
OneR	65%	60%	Average	Good
PART	100%	60%	Good	Good
Support Vector Machine(SVM)	100%	60%	Good	Good
K-Nearest Neighbor(KNN)	100%	60%	Good	Good

Table 5.15: Performance comparison of Molecular biology dataset.

decision is also “Average” because support vector classifier classified this molecular dataset with 57.21% accuracy where standard accuracy is 51.88%. So in this domain, the prediction of our meta-knowledge is 100% accurate. Also, it can be concluded that our meta-knowledge is almost 98% accurate for this test dataset.

Test 3:

For the test 3, we selected yeast dataset which is from Life area, has 1484 instances, 10 attributes, Attribute type is Nominal also it has no missing value and streaming data source. After converting the information into meta information, size of the dataset is “Large”, Dimension is “Low”, Attribute type is “Nominal”, streaming data is “No” and the missing data feature is “No”.

Algorithm Name	Accuracy (Weka)	Standard Accuracy (ZeroR)	Decision (Weka)	Decision (Meta Knowledge)
Naive Bayes	57.61%	31.20%	Average	Average
Decision Tree	50.34%	31.20%	Average	Bad
Random Tree	38.81%	31.20%	Average	Bad
Random Forest	49.12%	31.20%	Average	Average
Bagging	31.60%	31.20%	Average	Bad
Boosting (Adaboost)	40.70%	31.20%	Average	Average
OneR	18.87%	31.20%	Bad	Bad
PART	48.99%	31.20%	Average	Bad
Support Vector Machine(SVM)	59.70%	31.20%	Average	Average
K-Nearest Neighbor(KNN)	52.29%	31.20%	Average	Average

Table 5.16: Performance comparison of Yeast data set.

According to the Table 5.16, the decision of meta-knowledge for the naive bayes classifier is “Average” that means this naive bayes classifier can be able to classify this test dataset with more than standard accuracy but it will have less than 90% accuracy. In weka, the decision is also “Average” because naive bayes correctly classified 57.61% using ten-fold cross-validation. So in this domain, the prediction of our meta-knowledge is 100% accurate.

Test 4:

For the test case 4, we selected balloons dataset which is from Life area, has “Small” on the size of dataset feature, Dimension is “Low”, Attribute type is “Nominal”, streaming data is “No” and

the missing data feature is “No”.

Algorithm Name	Accuracy (Weka)	Standard Accuracy (ZeroR)	Decision (Weka)	Decision (Meta Knowledge)
Naive Bayes	100%	60%	Good	Good
Decision Tree	100%	60%	Good	Good
Random Tree	100%	60%	Good	Good
Random Forest	100%	60%	Good	Good
Bagging	90%	60%	Good	Average
Boosting (Adaboost)	100%	60%	Good	Good
OneR	65%	60%	Average	Good
PART	100%	60%	Good	Good
Support Vector Machine(SVM)	100%	60%	Good	Good
K-Nearest Neighbor(KNN)	100%	60%	Good	Good

Table 5.17: Performance comparison of Balloons dataset.

According to the Table 5.17, for the support vector algorithm, our meta-knowledge decision is “Good” that means this support vector classifier can be able to classify this test dataset with more than 90% accuracy. In weka, the decision is also “Good” because after ten-fold cross-validation the accuracy is 57.21% where standard accuracy is 51.88%. So in this domain, the prediction of meta-knowledge is 100% accurate.

Test 5:

For the test case 5, we have selected the Roth dataset which is from the Social area and the size of the dataset is “Small”, the dimension is “Low”, attribute type is “Numeric”, streaming data is “No” and the missing data feature is “No”.

Algorithm Name	Accuracy (Weka)	Standard Accuracy (ZeroR)	Decision (Weka)	Decision (Meta Knowledge)
Naive Bayes	80.30%	37.88%	Average	Good
Decision Tree	72.73%	37.88%	Average	Average
Random Tree	46.97%	37.88%	Average	Average
Random Forest	67.42%	37.88%	Average	Good
Bagging	37.88%	37.88%	Average	Average
Boosting (Adaboost)	43.18%	37.88%	Average	Good
OneR	38.64%	37.88%	Average	Average
PART	74.24%	37.88%	Average	Average
Support Vector Machine(SVM)	79.55%	37.88%	Average	Good
K-Nearest Neighbor(KNN)	62.12%	37.88%	Average	Good

Table 5.18: Performance comparison of Roth data set.

According to the Table 5.18, the decision of meta-knowledge for the naive bayes classifier is “Good” that means naive bayes classifier can be able to classify the dataset with more than 90% accuracy. But we got 80.30% accuracy using cross-validation where standard accuracy is 37.88%. So in this domain, our meta-knowledge has over-predicted the accuracy. In average, the prediction rate of accuracy is almost 98%.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis presented an overview of finding an optimal solution for extracting meta knowledge from meta data. Meta learning allows the system to find the regularities and patterns in training data or previous knowledge and extract the knowledge from them. A Meta knowledge can estimate the accuracy of any data set without running any evaluation. So, we can save our time and prevent the repetition process to select appropriate classifier for a data set. Our proposed meta knowledge can predict the performance of an algorithm for any dataset. We verified our proposed model with more than 98% accuracy rate.

6.2 Future Work

While we have achieved the primary objective of the research, our future work will be focused on working with more algorithms and more data. The use of more feature data would be investigated since that have an important influence on the extracting meta knowledge obtained at the end. Concerning the extracting meta knowledge, we can also expect to improve the model performance with more test data.

Bibliography

- [1] C. Lemke, M. Budka, and B. Gabrys, “Metalearning: a survey of trends and technologies,” *Artificial Intelligence Review*, vol. 44, no. 1, pp. 117–130, 2015. 1, 2, 20
- [2] A. M. Almana, M. S. Aksoy, S. M. Abu-Soud, J. S. Deogun, V. V. Raghavan, A. Sarkar, L. A. Kurgan, and K. J. Cios, “An overview of inductive learning algorithms,” 2014. 1
- [3] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, *Metalearning: Applications to data mining*. Springer Science & Business Media, 2008. 1, 2, 3, 38
- [4] C. Alippi and P. Braione, “Classification methods and inductive learning rules: what we may learn from theory,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 5, pp. 649–655, 2006. 1
- [5] R. Vilalta, C. Giraud-Carrier, P. Brazdil, and C. Soares, “Using meta-learning to support data mining,” *International Journal of Computer Science Applications*, vol. 1, 2004. 1
- [6] C. Castiello, G. Castellano, and A. M. Fanelli, “Meta-data: Characterization of input features for meta-learning,” in *Modeling Decisions for Artificial Intelligence*, V. Torra, Y. Narukawa, and S. Miyamoto, Eds. Springer Berlin Heidelberg, 2005. 3, 19, 20
- [7] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. 4, 6
- [8] M. S. Pervez and D. M. Farid, “Literature review of feature selection for mining tasks,” *International Journal of Computer Applications*, vol. 116, no. 21, 2015. 4
- [9] J. Brownlee, *Master Machine Learning Algorithms: discover how they work and implement them from scratch*. Jason Brownlee, 2016. 4, 5
- [10] D. Farid and M. Zahidur Rahman, “Anomaly network intrusion detection based on improved self adaptive bayesian algorithm,” *Journal of Computers*, vol. 5, pp. 23–31, 2010. 5
- [11] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, “Hybrid decision tree and naive bayes classifiers for multi-class classification tasks,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1937–1946, 2014. 6

- [12] D. M. Farid, L. Zhang, A. Hossain, C. M. Rahman, R. Strachan, G. Sexton, and K. Dahal, "An adaptive ensemble classifier for mining concept drifting data streams," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5895–5906, 2013. 6
- [13] M. Y. Arafat, S. Hoque, and D. M. Farid, "Cluster-based under-sampling with random forest for multi-class imbalanced classification," in *11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. IEEE, 2017, pp. 1–6. 6
- [14] F. Y. Tani, D. M. Farid, and M. Z. Rahman, "Ensemble of decision tree classifiers for mining web data streams," *International Journal of Applied Information Systems*, vol. 1, no. 2, pp. 30–36, 2012. 6
- [15] D. M. Farid, C. M. Rahman, N. Harbi, E. Bahri, and M. Z. Rahman, "Attacks classification in adaptive intrusion detection using decision tree," *International Conference on Computer Science (ICCS)*, 2010. 6
- [16] T. R. Patil and S. Sherekar, "Performance analysis of naive bayes and j48 classification algorithm for data classification," *International Journal of Computer Science and Applications*, vol. 6, no. 2, pp. 256–261, 2013. 6
- [17] F. Rayhan, S. Ahmed, A. Mahbub, M. Jani, S. Shatabda, D. M. Farid *et al.*, "Cusboost: Cluster-based under-sampling with boosting for imbalanced classification," *2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS-2017)*, 2017. 7
- [18] J. R. Quinlan *et al.*, "Bagging, boosting, and c4. 5," in *Innovative Applications of Artificial Intelligence Conference, Vol. 1*, 1996, pp. 725–730. 7
- [19] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996. 7
- [20] D. M. Farid, M. Z. Rahman, and C. M. Rahman, "An ensemble approach to classifier construction based on bootstrap aggregation," *International Journal of Computer Applications*, vol. 25, no. 5, pp. 30–34, 2011. 7
- [21] D. M. Farid, C. M. Rahman, and M. Z. Rahman, "Adaptive intrusion detection based on boosting and naive bayesian classifier," *International Journal of Computer Applications*, 2011. 7
- [22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997. 7

- [23] Y. Freund, R. E. Schapire *et al.*, “Experiments with a new boosting algorithm,” in *International Conference on Machine Learning*, vol. 96, 1996, pp. 148–156. 8
- [24] D. M. Farid, G. M. Maruf, and C. M. Rahman, “A new approach of boosting using decision tree classifier for classifying noisy data,” in *2nd International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 2013, pp. 1–4. 8
- [25] A. A. Afza, D. M. Farid, and C. M. Rahman, “A hybrid classifier using boosting, clustering, and naive bayesian classifier,” *World of Computer Science and Information Technology Journal*, vol. 1, pp. 105–109, 2011. 8
- [26] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 61–72, 2013. 8, 9
- [27] D. M. Farid, M. A. Al-Mamun, B. Manderick, and A. Nowe, “An adaptive rule-based classifier for mining big biological data,” *Expert Systems with Applications*, vol. 64, pp. 305–316, 2016. 8
- [28] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *17th International Conference on Pattern Recognition*, vol. 3. IEEE, 2004, pp. 32–36. 9
- [29] N. F. Haq, A. R. Onik, M. A. K. Hridoy, M. Rafni, F. M. Shah, and D. M. Farid, “Application of machine learning approaches in intrusion detection system: a survey,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 3, pp. 9–18, 2015. 9
- [30] M. S. Pervez and D. M. Farid, “Feature selection and intrusion classification in nsl-kdd cup 99 dataset employing svms,” in *8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. IEEE, 2014, pp. 1–6. 9
- [31] R. O. Duda, P. E. Hart, D. G. Stork *et al.*, “Pattern classification,” *International Journal of Computational Intelligence and Applications*, vol. 1, pp. 335–339, 2001. 9
- [32] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, “Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks,” *Computing Research Repository*, 2017. 11
- [33] D. Dheeru and E. Karra Taniskidou, “UCI machine learning repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml> 12
- [34] S. Sayad, “Model evaluation.” [Online]. Available: http://chem-eng.utoronto.ca/~datamining/dmc/model_evaluation.htm 19

Appendix A

Sample code

For the experiment, we used Weka java library with our own implementation. The partial code we used is provided here.

```
import weka.classifiers.Classifier;
import weka.classifiers.bayes.NaiveBayes;
import weka.classifiers.evaluation.Evaluation;
import weka.classifiers.functions.SMO;
import weka.classifiers.lazy.IBk;
import weka.classifiers.meta.AdaBoostM1;
import weka.classifiers.meta.Bagging;
import weka.classifiers.rules.DecisionTable;
import weka.classifiers.rules.OneR;
import weka.classifiers.rules.PART;
import weka.classifiers.trees.DecisionStump;
import weka.classifiers.trees.J48;
import weka.classifiers.trees.RandomForest;
import weka.classifiers.trees.RandomTree;
import weka.core.FastVector;
import weka.core.Instances;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.text.DecimalFormat;
```

```
import java.util.Random;
```

```
    public class WekaHelper {
static BufferedReader reader;
static Instances data,trainSet,testSet,newTrain;
public static String trainingCal(Classifier nb,Instances data,String filename)
throws Exception{
    final DecimalFormat df = new DecimalFormat("#.###");
    Evaluation eval = new Evaluation(data);
    eval.evaluateModel(nb,data);
    String str = filename+"",'+
    df.format(eval.pctCorrect())+",'+df.format(eval.weightedPrecision())+",'+
    +df.format(eval.weightedRecall())+",'+df.format(eval.weightedFMeasure())+'";
    return str;
}
```

```
    public static String percentageSplit(Classifier nb,Instances train, Instances test,String file-
name) throws Exception{
    final DecimalFormat df = new DecimalFormat("#.###");
    //test set
    Evaluation eval = new Evaluation(trainSet);
    eval.evaluateModel(nb,testSet);
    String str = filename+",'+df.format(eval.pctCorrect())+",'+df.format(eval.weightedPrecision())
+','+df.format(eval.weightedRecall())+",'+df.format(eval.weightedFMeasure())+'"; return str;
}
```

```
public static String crossValidation(Classifier nb,Instances data,String filename) throws Excep-
tion {
    final DecimalFormat df = new DecimalFormat("#.###");
    Evaluation eval = new Evaluation(data);
    eval.crossValidateModel(nb, data, 10, new Random(1));
```

```
    String str = filename+",'+df.format(eval.pctCorrect())+",'+df.format(eval.weightedPrecision())+",'+
    df.format(eval.weightedRecall())+",'+df.format(eval.weightedFMeasure())+'";
```

```
    return str;
}
public static void main(String[] args) throws Exception {
```

```

File folder = new File("dataset");
File[] listOfFiles = folder.listFiles();
for (int j = 0; j < listOfFiles.length; j++) {
    try {
        reader = new BufferedReader( new FileReader(listOfFiles[j].getPath())); data = new Instances(reader);

        int trainSize = (int) Math.round(data.numInstances()*0.70);
        int testSize = data.numInstances()-trainSize;

        int seed=1;
        Random rnd = new Random(seed);
        data.randomize(rnd);

        trainSet = new Instances (data, 0, trainSize);
        testSet = new Instances (data, trainSize,testSize);

        reader.close();
    } catch (IOException e) {
        e.printStackTrace();
    }
    data.setClassIndex(data.numAttributes() - 1);
    trainSet.setClassIndex(trainSet.numAttributes() - 1);
    testSet.setClassIndex(testSet.numAttributes() - 1);
    int k1;
    k1 = (int) Math.round(data.numInstances()*0.05);
    Classifier[] models = {
        new AdaBoostM1(),
        new Bagging(),
        new IBk(k1),
        new J48(),
        new NaiveBayes(),
        new OneR(),
        new PART(),
        new RandomForest(),
        new RandomTree(),
        new SMO(),
    };

```

```

String[] classifierName = {"AdaBoostM1","Bagging","IBK","J48","NaiveBayes","OneR",
"PART","RandomForest","RandomTree","SMO"};
String str;
System.out.println("Running ...");
for(int i = 0; i < models.length; i++) {
FileWriter trainFW = new FileWriter("result+classifierName[i]+train.csv", true);
FileWriter crossFW = new FileWriter("result+classifierName[i]+cross.csv", true);
FileWriter testFW = new FileWriter("result+classifierName[i]+test.csv", true);

    try {
models[i].buildClassifier(data);
str = trainingCal(models[i],data,listOfFiles[j].getName());
// training set evaluation writing in the table.
trainFW.write(str);
str = crossValidation(models[i], data,listOfFiles[j].getName());
// Cross validation set evaluation writing in the table.
crossFW.write(str);
// percentage split set
models[i].buildClassifier(trainSet);
str = percentageSplit(models[i], trainSet, testSet,listOfFiles[j].getName());
// testing on the percentage split evaluation writing in the table.
testFW.write(str);

    } catch (Exception e2) {}

    trainFW.close();
testFW.close();
crossFW.close();
}
} } }

```