Parameter Tuning Method Analytics (PaTM) for Different Datasets Using Classification Models

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science

Submitted By

Lata Dubey

(Roll No. 801532028)

Under the supervision of:

Dr. Seema Bawa
Professor
Dr. Anju Bala
Assistant Professor

Thapar University, Patiala

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2017

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Parameter Tuning Method Analytics (PaTM) for Different Datasets Using Classification Models*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Seema Bawa* and co-guide *Dr. Anju Bala* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Lata Dubey

(Lata Dubey)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Anju Bala

(Dr. Anju Bala)
Assistant Professor
Computer Science and Engineering Department

Seema Bawa
14/07/2017

(Dr. Seema Bawa)
Professor
Computer Science and Engineering Department

Abstract

In machine learning many classification models, have a range of parameters that may strongly affect the predictive performance of the model induced by them. Hence, it is recommended to define the values of these parameters using the optimization techniques. While these techniques usually converge to a good set of values, they typically have a high computational cost, because many candidate sets of values are evaluated during the optimization process. It is often not clear whether this will be a given results in parameter settings that are significantly better than the default settings. When training time is limited it may help to know when this parameter should definitely be tuned. Hence, in this thesis learning method has been used to predict when optimization techniques are expected to lead models whose predictive performance is better than those obtained by using default parameter settings. The parameter tuning method has also been utilized to improve the performance of classification models and reducing the error rate along with overall computational costs. We evaluate the proposed method on different datasets by selecting six datasets. The performance of parameter tuning framework is being evaluated and results show that the parameter tuning framework outperforms than other existing techniques.

Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisor **Dr. Seema Bawa,** Professor, Computer Science & Engineering Department and coguide **Dr. Anju Bala,** Assistant Professor, Computer Science & Engineering Department. They have been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of them. I also thank my supervisor and coguide for their time, patience, discussions and valuable comments. Their exultation and optimism made this experience both rewarding and enjoyable. I am truly grateful to her for extending her total co-operation and understanding whenever, I needed help and guidance from her. I am also heartily thankful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department and **Dr. Ashutosh Mishra**, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to Nishtha Hooda, PHD Scholar, the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Lata Dubey

Lata Dubey
(801532028)

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This chapter discussed about the optimization, parameter tuning, and various algorithms of machine learning to improve the performance of classification models.

1.1 Machine Learning

Machine learning is one of the intense growing levels of computer science, with far-arriving applications. To solve the computer problems, we have required an algorithm. Algorithm is a chain of orders should be done by changing the input data from output. The input data is the set of numbers and the output is pursued by them [1]. For the similar purpose, there are different machine learning algorithms and we get interested to discover the most capable one for the least amount of memory or instruction or both are required. Machine learning is used to optimize the performance criteria of the utilize data or past experience data for computing purpose through the computers. Machine learning is extremely related with the computational data, which is also focused on the forecasting during use of computers. It provides a strong link in mathematical adaptation that provides different approaches, principles and application domains in different areas. In the pasture of data analytics, we have used machine learning method to prepare the difficult models and algorithms that provides them to the prediction, which is known as the approximate analytics in commercial use [1, 2].

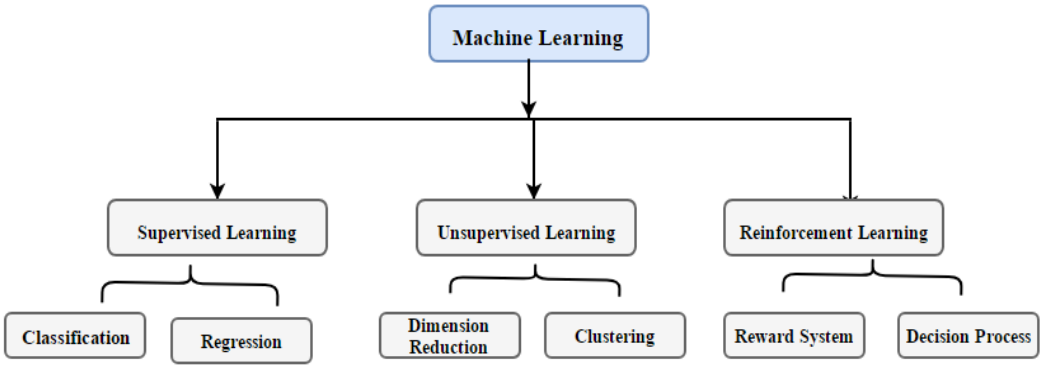

Fig 1.1: Types of Machine Learning

Machine learning method is used to huge databases is also known as data mining. Machine learning is a logical training inspects that the computational principle of learning, as a result, this is a essential inattentive of the possibility that we are just desirous on how we learn people and creatures. How the people and creatures learn. As a logical train, the machine inspects the computational base of to learn; As a result, it is fundamental that we are curious about this possibility regardless of how we learn people and creatures. There are different types of machine learning shown in (Fig 1.1).

1.1.1 Types of Machine Learning

Learning is an extremely broad area. There are numerous subfields to manage with dissimilar category of learning in area of machine learning. We give an extraordinary scientific categorization of learning standards, which is to give some viewpoint zone of the machine learning [1]. The following are the two most important classifications of machine learning act [2].

A. Supervised Learning

A supervised machine learning algorithm split downward the training information and produced a derived capacity, which can utilized for mapping new cases. Supervised learning as regression (for persistent yields) and order (for discrete yields) is a critical constituent of insights and machine learning [3]. In the supervised learning you have the input variables (X) and the output variables (y) and then apply the algorithm through input mapping function to the output.

$$Y=f(X)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ (Eq.1)}$$

The goal of the mapping function is so well estimated that when you have new input data (x), you can guess the output variable (y) for that data [3]. This is called supervised learning because the learning process of the algorithms with training datasets can be considered as a supervisor. Supervised learning is divided into two expanded categories; these are classification and regression [3, 4].

i. Regression: Regression method is use to identify the dependent variable that is associated with the independent variable, to find out the forms of relationships [4]. It attempts to clearly demonstrate the connection between input or independent variables and outputs, usually in the form of parametric equations in which the parameter is estimated from the data [5]. Regression analysis is widely used for forecasting and forecasting, where it is used to overlap sufficiently with the field of machine learning. Regression models describe in three different categories: linear, polynomial and logistics regression [5, 4].

a) **Linear regression:** A linear regression method has been explained as a modeling of connection between the scalar dependent variables and the independent variables. There is only one independent variable, we also known as simple linear regression model. And if the independent variables are more than one, is known as multivariate linear regression [7]. Here, the best fit line is called as regression line and described with a linear equation:

$$Y= a *X + b \quad \text{(Eq.2)}$$

b) **Polynomial Regression:** Any regression equation will be polynomial regression when the power of independent variable is greater than one. The given below equation express a polynomial equation [7].

$$Y=a+b*x^2 \quad \text{(Eq.3)}$$

The best fit line in the regression technique is not a straight line. But it is a kind of curve which fits into data points.

c) **Logistic Regression:** We have used logistic regression method to discover the probability of any event success and failure. We are using logistic regression method when dependent variables are binary (0/ 1) (True/ False) (Yes/ No) in nature [7].

ii. Classification: Classification is used in Machine Learning, Data Mining, and Statistics because there is an issue of recognizing a group of classification categories, in which a new perception is incorporated, and the observational information depends upon the training data set and whose classes are also known as membership. We can partition the classification in two sections [6].

a) **Binary Classification:** When we have categorized the given data into two distinct classes is called as binary classification.

b) **Multiclass Classification:** The volume of squares is greater than two is also known as multiclass classification.

B. **Unsupervised learning**

Unsupervised learning is a method of learning where only input data (X) available but no related output variables. To learn more about the data, the purpose of unsupervised learning is to understand the underlying structure of data. These are called unsupervised learning because there is no right answer in contrast to the above supervised education and there is no supervisor [3]. The algorithm is left in the data for its own devices to find and present an interesting structure. Unsupervised learning can categorize in two different problems they are as follow:

i. Clustering: Clustering is also a technique of unsupervised learning. In the unsupervised learning, clustering is one of the very frequent methods on the statistical data analysis that is also used into the different areas of real world [9]. Clustering models are different from supervised models in the result of known results, i.e., there is no target attribute. On the other hand, clustering models have been created using optimization criteria that support high interval cluster and low inter cluster similarity [3].

ii. Association: There is a problem learning an association rule, where you want to find the rules that describe a large part of your data, such as those buying X, also buy Y [3].The following are the few popular examples of unsupervised learning. These are:

i. K-means for clustering problems.

ii. Apriori algorithm for association rule learning problems.

1.2 Optimization

Optimization is the selection of the best element from some set of available options. In the problem of optimization, the actual work can be maximized or minimized by selecting the input values from within a real set and calculating the value of the function [8]. Generally, optimization involves the "best available" value of some objective functions given to a certain domain (or input), including a variety of objective functions and different types of domains [4]. There is a scope of improvement of every phase of machine learning. There are some methods of optimization through which we can improve the performance of models. These are as following:

1.2.1 Ensembling

Ensembling is a learning procedure by many models, for example, classifiers or professionals, emerge strategically and join to tackle the issue of a specific computational knowledge. Learning is mainly useful for improving the performance of the classifications, forecasting and the function approximation, etc. of the given models or to reduce the possibility of the poor unwanted collection of data [10].

1.2.2 Feature Selection

Feature selection is also called variable selection or attributes selection. This is the automatic selection of attributes in your data (such as column in structural data) that are most relevant to the forecast modeling problem you are working on. Feature selection is different from dimensional reduction. Both ways seek to reduce the number of properties in the dataset, but one-dimensional reduction method does this by creating new combinations of properties, where the features are included in the selection methods and without changing the attributes in the data, is removed [10].

A very few research work has been done in the area of parameter tuning hence, various parameter tuning approaches have been explored.

1.3 Parameter Tuning

Parameter tuning method is an extremely significant method of machine learning that is useful for the evaluated the model performance therefore the performance of the model has been improving. Parameter tuning is used to boost the performance of machine learning models [12]. In classification models, have a range of parameters that strongly affect the predictive performance of the models. Choosing the parameter tuning method to improve the performance of the classification models and reducing the error rate and the overall computational costs [12, 11].

1.3.1 K-Fold Validation

In the optimization technique K -Fold validation process is one of the very imperative fractions on the whole analysis. This method is defined in an extremely crisp way, using the machine learning algorithms. The perception of examining stability of models, predicting the target values is one of the main concepts. In our case, we will examine 10 times the validation in the various data mining and machine learning models with 10 diverse times for dissimilar datasets, each point in time dissimilar training and test data sets. The selected parts the accuracy for the check stability, if our models accuracy is provide consistent and the given results are suitable, then we has concluded that model really predicts the target values of a good model [7]. K-Fold validation method has been applied to increase the performance of machine learning models.

1.4 Evaluation Metrics

Evaluation metrics plays significant role in machine learning. Which are employed to measure the learning algorithms. The commonly used metric for these intents is the accuracy. Yet, on an imbalanced data set, accuracy is not an appropriate metric, since the positive course of instruction has little effect on classification rate (accuracy) as compared to negative class hence; other evaluation parameter has been used to enhance performance of classification algorithms [5].

Table 1.1: Confusion Matrix [5]

	P' (Predicted)	N' (Prdicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

There are the many metric used for estimating performance of classifiers. There are some of them defined below [6].

1.4.1 Accuracy

It is a ratio of correct predictions into number of instances evaluated.

$$\text{Accuracy} = {}^{(TP + TN)}\big/_{(TP + FN + TN + FP)} \text{…….. (Eq. 4)}$$

Where,

TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative respectively.

1.4.2 ROC

ROC curve is used to design the false positive rate on the X-axis and the true positive rate on the Y-axis.

$$\text{Y-axis => TPR} = {}^{TP}\big/_{(TP + FN)} \text{…….. (Eq. 5)}$$

$$\text{X-axis => FPR} = {}^{FP}\big/_{(TN + FP)} \text{…….. (Eq. 6)}$$

1.5 Classification Models

Here, different machine learning algorithms to improve the performance of the classifiers. The some of the classifiers have been explained in following:

1.5.1 SVM: Support Vector Machine (SVM) is a very useable model in supervised learning that is broadly used into pattern recognition that is used into regression models and also used in classification modeling Assume there are a few cases for

train the SVM model will attempts to classify in a particular manner respectively keen on the large amount of a high accuracy for the classification models [7]. SVM is a clearly standout amongst the most usually utilized kernel learning algorithm that performs robust nonlinear classifications of sample using on kernel tricks [14].

1.5.2 Decision Tree: Generally, decision tree learning utilized the decision trees as a predicted models perception around a thing (spoken to in the branches) to decisions about the thing's objective esteem (spoken to in the clears out). It is one of the predictive modeling approaches utilized as a part of insights, information mining and Machine Learning [7, 13].

1.5.3 Logistic Regression: Logistic regression is a binary classification algorithm. It assumes that the input variables are numeric and Gaussian (bell curve) is distributed. This last point should not be right because your data is not Gaussian, then a regressive regression can still get good results [7]. Estimate the logistic regression models using the lrm function from the Design package for the R statistical programming language [15].

1.5.4 Bagging: Bagging (Bootstrap aggregation) is a voting system where construct models are found out in light of various adaptations of learning informational indexes that are created by bootstrapping (bootstrap sampling). Using an unstable learning algorithm would be to use bagging (e. g., neural networks or decision trees), result in largely different classifiers when small changes in the learning set. There is a proposal to detect a novel intrusion based on the wearable method of Machine Learning. As a base class, the bagging method of the garb with REP tree is used to implement intrusion detection system [15].

1.5.5 Stacking: Stacking is a technique for joining independent base models, models learned among various learning algorithms, for example, the k-nearest neighbor technique, decision trees, Naive Bayes, and so on. It is also known as stacked speculation. Base models aren’t joined with a settled plan, for example, voting,

yet slightly an extra model are also known as meta-models that are found out and utilized the consolidating base models. This methodology has 2 stages. To start with, we make the meta-learning educational record via the desires of the base models. With the use of meta-learning sets we are taking in the meta-display, which can merge desires of the base models hooked on the last estimate [16].

1.5.6 AdaBoost: Boosting contains an entire group of comparable technique that, similarly since bagging, make use of voting to merge the figures for base models well-read by a solo learning calculation. The contrast among the more than one methodology during bagging integrally of built base models that is absent to risk, whereas boosting attempts to create reciprocal base model by the learning resulting models, considering the missteps of the precedent models. This methodology begins by learning the reputable initial point models that lying on the whole learning set with similarly weighted cases [17].

1.5.7 Random Forest: Normally, random forest is one of most popular machine learning algorithms. Random Forest is nothing but a group of many simple decision trees and all these trees are able to predict the outcome for any input. These trees are able to predict in which class a particular input belongs to if our problem is of classification and if problem is of regression problem these trees are able to predict a continuous number. In this, case of classification each tree in random forest votes for a particular class and the class which have most votes is given as output for that particular input on the other hand in regression output of every tree is averaged to obtain the output for that particular input. Random Forest can be seen as ensemble of many simple decision trees. Ensembling of many decision trees in random forest have shown dramatic improvement in performance of model. Random Forest is also able to overcome the issue of over-fitting which is one of biggest problem in single decision tree [18].

Chapter 2: Literature Review

In this chapter, the survey on Machine Learning, classification and parameter tuning methods for various classification models have been performed. The following is the summary of the review performed by various research fellows.

2.1 Machine Learning

Simon et al. [19] determined about machine learning that is used in different field, defend stable by general intension by a comparable assessment method. The common goal of this method to get better performance on a few particular responsibilities and common method the data includes regularity and exploitation in training dataset. Mainly evaluation performed by logically, for the purpose of this kind of method of learning points to the performance on the different test sets that generated in one or other prudent slots, which gives the higher quality performance on the given test sets without any trial.

Najafabadi et al. [22] discussed about machine learning that is generalized one of the learned patterns for given input data and usage on future unnoticed data. The data present a major concussion upon the concert of given machine learning models on the goodness of the data. The poor data represents a probability to decrease the performance of the superior, complicated machine learner, while a high quality data evaluation can be an elevated performance for the comparatively straightforward machine learning models.

Rasmussen et al. [20] considered the pre occupation of stochastic procedure and how we will be usage on different machine learning algorithms. They had conferred simple equations to include training data sets and checkups. To interpreted the practicable benefits of Gaussian process that are ending in conclusion and see ongoing trends in GP's work.

Woon et al. [21] focused on the uses of the Machine learning algorithm that is differentiate with diverse types of fractional discharge, which have been closely related to

insulation stoppage. Measures used for acoustic emission sensors are used to train and test the sound classification algorithms. In this study, the ability of higher classification was obtained through training and testing datasets composed inferior the same classification conditions. However, the accuracy of classification algorithms was very low in different circumstances. Different experiments are using in the most recent classification techniques were shown, major improvements are shown in performance of classification accuracy.

Najafabadi et al. [22] discussed about some important dimensions of deep learning researches, in which include some explicit challenges presented by the Big Data Analytics, further research are required, together with emerge data, high-dimensionality data, scalability of the models, and the distributed computing techniques. There are few approaches to execute machine learning process, but these two are mainly used in this approach they are as: supervised and unsupervised learning.

2.1.1 Supervised learning

Kotsiantis et al. [23] described the main purpose of the supervised education is to represent in the context of predictive features; a summarizing model of partition of class labels is to be created. As a result, classifier is used to hand over the class labels in the test examples, there is the values of the launcher facet are also known, but the values of the class labels are unknown. This paper is used distinct supervised machine learning classification techniques.

Chitra et al. [24] proposed supervisor classifiers known as SVM (vector vector machine) is adoptive for predicting cardiovascular disease using the patient's medical records in the early stages and the result collate with the identified supervised classifiers support vector machine (SVM). The brain record instructions are classified into the cascaded neural network (CNN) classifier. The proposed system will be providing assistance to physicians in diagnosing the disease more efficiently.

Kotsiantis et al. [25] described about the multiple classification algorithms and the neoteric efforts to improve the performance of classification algorithms. Main objective

of supervised learning is to design a brief model of the distribution class labels in terms of predicted features. As a results, classifiers is used to hand over the class label in test examples, the values of launcher features also known, but the values of the class labels is unknown.

Yang et al. [26] focused on numerous method of regularization that is mainly used to obtain structural information of the given data. The methods of several regularization are mainly used in the laplacian graph, and ordinarily it is unsupervised learning. To solve the difficulties of supervised learning, we have proposed a supervised method to calculate laplacian graph using the hellinger distance to scale up the equality of the sample dataset. Hellinger distance can be grant an extensive evaluation that is related with samples of four aspects, which have been provide the similarity, density, dimension and direction. The conventional linear model or support vector machine desires to improve when we will deal with datasets with many attributes or multi-sources. So, that they have recently proposed many kernel learning by adaptable and flexible way. In this proposed classification models we have used manifold regularization to include structural information to disrupt many kernel classifiers, which hope to achieve classifier which reduces classification errors and understands the structural information about the dataset. So that we have organized the whole data should be the proper view. In this experiment, they can use the UCI repository dataset to exhibit the classification performance of proposed model, and they can use a synthetic dataset to validate the manifold regularization that can obtained the parts of structural information

Figueroa et al. [27] proposed an innovative perspective based on an ensembling methods of classifiers. They had used this method to combine syntactic and semantic features so that we will effectively detect user severity. In which diverse experiments setting shows that obligation of linguistically motivated ensembling perspective, during reducing the position variance of a single classifier in the user intentions.

Sujatha et al. [28] provided the method of manifolds on data mining and machine learning classification techniques used for innovatory data based applications. Classification model can use to find the process that used to assign the data into number

of classed according to dissimilar constrains. In this paper, includes several most important types of classification algorithms like genetic algorithm C4.5, Naive bayes, Support Vector Machine, KNN, Decision Tree, and CART.

2.1.2 Unsupervised Learning

Huang et al. [29] used the supervised, unsupervised, and semi-supervised, ELMs (Extreme learning machines) that can actually be insert into an integrated framework. This random feature supplies a new probability to realize the mechanism of mapping, it is an important conception in ELM theory. Empirical study on detailed data show that the suggested algorithms are competing with state-of-the-art semi supervision or unified learning algorithms in terms of accuracy and efficiency.

Pang et al. [30] proposed a novel Coupled Unsupervised Feature Selection structure (CUFS) that is to filter out the noisy and duplicate features for subsequent outlier detection in classified data set. CUFS had been quantified the contingency of features by learning or integrating both the quality coupling and the feature coupling. The value of feature coupling capture the internal data attributes and separate characteristics related to the duplicate or unnecessary features. CUFS were further instantiate into a parameters-free Dense Sub graph-based on Feature Selection method, known as DSFS. We had proved that the DSFS retains a adjacency feature subset to the optimal feature subset.

Ferreira et al. [31] had used the entropy-based data selection as a substitute of random equiprobable sampling ahead of training models, considerable improvements are achieved in parameter convergence, accuracy and the popularization capability. In addition, the model evaluation metrics demonstrate a smaller amount of variance, consequently allowing quicker junction when the multiple modeling tests have executed. These features are being used by experimentally and placed by the results of a wide neural network predictions modeling experiments, where the identification of pairs of models was a data set used to tune the model parameters values. Unlike the most active learning and example selection processes, this method is not repeated, does not rely on the existing model, and the specific modeling techniques are not required.

Zhang et al. [32] presented the performance evaluation usability for the six distinctive unsupervised feature selection algorithms on the facet of accuracy, time cost, and the hyper-parameters. Here, an ordinary scrap based structure with the selection of middling parameters that have adopted toward, underline the divergence between the algorithms. The given experiments are verified that the sparse coding can be attaining the steady concert across the diverse datasets. Furthermore, the random patches through soft threshold functions and K-means combining with the triangle coding attain similar act with the sparse coding method, still more rapidly and easier to train, the results proposes that they are high-quality choices to construct an application system in exercises.

Asif et al. [33] had used a robust and scalable method using n-SVR to hold the difficulty of speed prediction of the huge heterogeneous road networks. The conventional performance of the dataset measures such as mean absolute percentage error (MAPE) and the root mean square error (RMSE) offer a small vision into the dimensional and temporal aspects of prediction methods on huge networks. This deficiency can a grave huddle in impressive exertion of a prediction models for a direction supervision, excess avoiders, aggressive traffic assignment and additional applications. We have proposed unsupervised learning approach with k-means clustering, self organizing maps (SOM), and principal component analysis (PCA), to reduce the deficiency. They had evaluated the effectiveness of developed methods by evaluating the spatial and temporary characteristics of the predicted performance of the proposed variable window-SVR method.

2.2 Applications of Machine Learning

Machine learning is one of the most exciting technologies that will ever be realized. As it is clear by name, it gives computers that make it like humans: learning ability. Machine learning is being actively used today, possibly more than one is expected in many places, probably use dozens of times without knowing a learning algorithm [6]. There are various domains where machine learning is used like; Medical Informatics, Big Data Analytics,

Data Mining, Researches. The following are the some applications of machine learning shown in Table 2.1.

Table 2.1: Machine Learning Applications

Sr.no	Paper Title	Paper Description	References
1.	An Application of Machine Learning to Anomaly Detection	The paper presents the machine learning approach for anomaly detection. System creates user based profile on command sequences; compare the current input sequence profiles using one equality measure. Our empirical results show that this is a great way to separate the legitimate user from an intruder.	[7].
2.	Marine Life Airborne Observation using HOG and SVM Classifier	The method is based on HOG (Histogram of Oriented Gradients) features extraction and SVM classification process. In which, classification using numerical HD photography like birds, marine, mammals, and other human race objects (garbage).	[8].
3.	Understanding Machine Learning: From Theory to Algorithms	Some special applications involved in machine learning such as: Text Categorization, Spam Detection, Speech recognition programs	[9].
4	Applications of Machine Learning and Rule Induction	Machine learning is a diverse area, which is organized simultaneously by identical goals and similar evaluation methods. The general purpose is to improve performance on some tasks, and to find regularity in training data.	[10].
5	Deep Learning for Health Informatics	The primarily applications of deep learning in the field of translated bioinformatics, Medical imaging, Extensive sensing, Medical informatics, and Public Health	[11].

There are numerous applications of machine learning. Here is a list of a few of them [19].

i. **Weather forecast**: Machine learning has applied to software's that forecasts weather so that the superiority can be enhanced.

ii. **Malware stop/Anti-virus:** With growing amount of malicious records every day, that is receiving impossible for the humans and many other security solutions to keep up hence, machine learning and deep learning are important. Machine learning is used to train the anti-virus software so that it can predict the improved results.

iii. **Anti-spam:** A machine learning algorithm provides the spam filtration algorithms to give the better distinguish spam emails from the anti-spam mails.

iv. **Face detection/Face recognition:** Machine learning can be used in many different devices like mobile, cameras, laptops, etc. it is also used for face detection and recognition methods. For instance, camera clicks a photo automatically when anyone smiles much more accurately now because of advancements in machine learning algorithms.

v. **Speech recognition:** Speech recognition systems have improved significantly because of machine learning. For example, Google.

vi. **Genetics:** Clustering algorithms in machine learning can be used to find genes that are associated with a particular disease. For instance, Medecision, a health management company, used a machine learning platform to gain a better understanding of diabetic patients who are at risk.

There are numerous other applications such as image classification, smart cars, increase cyber security and many more.

2.3 Performance Measures for Classification Algorithms

Li et al. [34] defined that the SVM models as a weak learning for the AdaBoost model, AdaBoost focused on the design of an algorithm called SVM. To achieve a group of influential SVM weak learners, this algorithm is used to adjust kernel parameters in SVM, rather than using a fixed optimum. To achieve a set of effective SVM weak

learners, this algorithm adjusts the kernel parameter in SVM rather than using a fixed one optimal. Compared to the current AdaBoost methods and AdaBoostSVM has the advantages of easy model selection and better generalization performance. This provides a possible way of handling more suitable problems in AdaBoost. An advanced version called AdaBoostSVM has been developed further to deal with the duality of accuracy / diversity in the Boosting Method. By implementing some parameter adjustment strategies, these SVMs see the distribution of accuracy and diversity on poor learners to gain a good balance. To do the best of our knowledge, a system that can easily and clearly balance. Experimental results showed that both proposed algorithms gain better generalization performance compared to AdaBoost using other types of weaker learners. Benefits of balance between accuracy and diversity, Diverse AdaBoostSVM achieve best performance. In addition, using unbalanced data sets revealed that AdaBoostSVM has performed better than SVM.

Assareh et al. [35] investigated the efficacy of implementing statistical and information theoretical strategies in collaboration with AdaBoost in order to improve its performance. The results show that the performance of AdaBoost is less sensitive to parameter learning settings when risk frequencies are less, which can be explained in relation to data fragmentation phenomena. Apart from this, depending on the model of interaction between risks SNP, different criteria can achieve excellence in the second phase.

Hashi et al. [36] focuses on to diagnosis diabetes disease as it is a great threat to human life worldwide. The system uses the Decision Tree and K-Nearest Neighbor (KNN) Algorithms as supervised classification model. Finally, the proposed system calculates and compares the accuracy of C4.5 and KNN and the experimental result demonstrates that the C4.5 provides better accuracy for diagnosis diabetes.

Joshi et al. [37] focused on a comparative study of different classifiers (Decision trees) when the ensemble learning technique called bagging is used. We perform classification on various datasets firstly by using a single classifier and then by bagging method, using

the same base classifier. It is observed that when they use a single classifier rather than an ensemble, the classification error further increases. Training data subsets at random prepared - with the alternate of the overall training dataset, so usually the new training set includes some duplicate and some defaults data compared to the original training dataset. Therefore, all the training data subset is used to train a different classifier of the same type.

Mantovani et al. [38] focused on how to determine hyper-parameter optimization procedure for sensitive decision Trees. In this paper the researchers are using four different tuning techniques were detected to adjust the J48-Diesion tree algorithm for evaluated hyper-parameters. In this method they have used total 102 heterogeneous datasets inspect the tuning effect on the motivated model.

Wang et al. [39] proposed a new way of spark's tuning configuration is proposed on the basis of the machine tuning process, more effective spark tuning process, which is composed by the binary classification and multi-classification. This method can be used to auto-tune the spark configuration parameters. In addition, many common machine learning algorithms are detected based on the proposed method, and experimental results show that decision tree model (C5.0) is the best model based on accurate and computational effectiveness.

2.4 Parameter Tuning Analysis

Mantovani et al. [40] predicted the models of meta-learning approach that expected from optimization techniques, whose predicted performance is achieved using default parameter settings. Therefore, we can choose optimization techniques only if they are expected to improve performance, and hope to reduce the total computational cost. They had evaluated these meta-learning techniques on more than one hundred data sets. Experimental results show that precise estimates can be made when using optimization techniques rather than default values suggested by some machine learning libraries.

Ding et al. [41] had defined the SVM algorithm based on PSO (particle swarm optimization). The author had been described the three classification methods to classify a benchmark spectral image, like SVM, ML (maximum-likelihood) and K-nn (K-nearest neighbor), in which the performance of SVM is compared with two different traditional classifiers (ML, K-nn). Therefore the study indicates that the performance of the SVM classification accuracy is better than the other two classifiers. In which an SVM algorithm is based on PSO to improve the classification accuracy compared with the original SVM classification.

Novakovic et al. [42] presented the kernel parameter to improve the accuracy of the classification and finds the best performance on the basis of the linear kernel. The biggest benefit of the parameter optimization had those kernel functions with the smaller accuracy of kernels. The time taken to build a model is very high with C-SVC (C-support vector classification) and polynomial kernel, compare with the other kernels.

Eiben et al. [43] presented a conceptual framework for parameter tuning, they provide a survey of tuning methods and discuss related technical issues. This framework is based on the three-level hierarchy of the problem, an evolutionary algorithm (EA) and a tuner. Apart from this, we separate examples of examples, parameters and EA performance measures as major cases and discuss how tuning can be guided to perform and / or strengthen algorithms. For the part of the survey, we set various taxonomies to classify tuning methods and to review the existing work. Finally, they explained in detail that the method of tuning can improve through well-funded experimental comparisons and algorithmic analysis.

Gao et al. [44] applied the PSO, SVM, and Genetic algorithm to improve the accuracy and reduce computational circulation based on a heuristic algorithm. The proposed technique is used to integrate the GA operators into PSO.

Chiu et al. [45] focused a method to find the impact of various parameters such as attribute, instances, and classes on clusters, accuracy, and diversity in a real world

problem using ensemble classifier. The primary aim was to find the link between the parameter, accuracy, and diversity, and the secondary goal was to find any relationship between the number of clusters in ensemble classifier and data variable.

Molina et al. [46] used the classification models which can help to decide how the default parameter tuned in order to increase the accuracy and ROC of the classifier when we using the different type of datasets.

Fei et al. [47] presented the number of support vector with the selected support vector also has an effect on classification performance of SVM. The main aim of this paper is to select the support vector and feature subset simultaneously based on genetic algorithm, and constantly to search the best penalty parameter C and Kernel function parameters. It would be interesting to extend the proposed approach which can apply to multiple classification tasks such as text classification and the other way to using the proposed approach with other kernel functions like the poly kernel and linear kernel to construct the several experiments.

Sherin et al. [48] presented the quality of the classifier depends on various factors, in which SVM parameter tuning is very compulsory A lot of efforts have been made for the important automatic kernel Choice and parameter optimization, meta-compatibility Algorithms such as genetic algorithms (G.A.) and particle swarm Optimization (PSO) This paper selects the option SVM Parameter, Kernel, and kernel parameter optimization by using the bat algorithm to highlights the results.

2.5 Research Gaps

This section tells about the gaps encountered during the research by reviewing the already existing literature in performance evaluation parameters in machine learning classifiers.

i. Improve the accuracy of the classification models and finds the best performance on the basis of the parameter values using the large number of datasets [42].

ii. Increase the performance of different parameter evaluation parameters of the classifier when using the different type of datasets [46].

iii. Scalability and reusability is a most important challenging problem in the field of machine learning. Traditional tools are not able to handle large datasets [43].

iv. Timeliness is another problem for large datasets in machine learning. As the size of datasets is increases, analyzing time is also increases [40].

2.6 Problem Formulation

The performance evaluation method is a crucial problem in the machine learning, which has become an emerging research area in recent years. Classification algorithms afflicted by the high dimensionality problem for a dataset would see strong overall accuracy but very low performance on the positive class. In this method analyzing an evolutionary algorithms by studying how the performance depends on the problem and how its performance varies when we have executing independent repetitions of its runs. Therefore, tuning method can improve the performance by facilitating well funded experimental comparisons and algorithm analysis. Performance evaluation problem occurs in many real-world areas like oil spills detection from satellite images, fraud detection, anomaly detection, medical diagnosis, identifying fraudulent credit card transactions etc. The data may be composed with large percentage of negative samples and less percentage of positive samples. Hence, researchers have given more focus on problem of evolutionary algorithms therefore the performance of the classification models will be improved.

2.7 Objectives

To study and analyze existing algorithms, methods, techniques and models for class imbalance.

i) To study and explore existing state-of-the-art machine learning classification algorithms, methods, techniques and models.

ii) To analyze an algorithms by studying how the model performance depends on its parameter values.

iii) To propose a framework for an efficient parameter tuning method.

iv) To evaluate the performance of proposed framework using various parameters like accuracy and ROC.

v) To check the robustness of our selected model or proposed framework using cross validation technique.

Chapter 3: Proposed Framework Parameter Tuning Method Analytics (PaTM)

To create the experiments more interesting, firstly we have collected the real world datasets. In the first phase, we have collected the binary and multivariate datasets, which have different attributes, different instances, and different class. After selecting the data, the second phase is data preprocessing. The real world data is generally incomplete, noisy and inconsistent. We will need to consider how will use the data. There are three common data preprocessing steps are formatting, cleaning, and sampling. Data formatting: The data we have selected may not be in a format. Therefore, in which the datasets must be converted into .CSV files which are required by R interface. Data cleaning: Cleaning data is used to the removal or fixing of missing data. Data sampling depends on two parameters – percentage and bias. Below Figure 3.1 shows the workflow of Parameter tuning method.

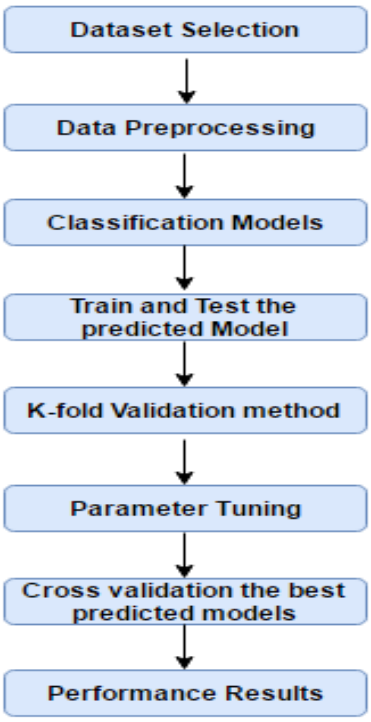

Fig 3.1: Workflow of PaTM

We have performed the performance evaluation on parameter tuning method for different classification algorithms. In which the classification includes with various classifiers such as SVM, Decision tree, Logistic regression, Bagging, Stacking, AdaBoost, and Random Forest. The classification models apply the different datasets and change the parameter values of the models. The comparison phase we focus on results achieved by above 7 classification algorithms on each datasets performance evaluation metrics: Accuracy, Error rate, and ROC and we have compare the obtained results to analyze the best classification algorithm for each dataset.

3.1 PaTM Framework

Evolution of classification algorithms have extracted great interest in recent years as it has been shown by several studies that, both theoretically and empirically, they can outperform single classifiers or multi classifiers to elaborate on how the tuning method can improve the methodology and classifiers by experimental comparisons and model analysis Since it is highly unlikely to train the perfect classifier that makes the well known results of the evolutionary models. The following are the important factors of parameter tuning method as shown in Fig (3.2) are discussed below-

3.1.1 Data Selection

Dataset selection is used collect data towards the appropriate data type and source decision, and additionally point to the appropriate equipment. Data collection occurs before real time in relation to data collection. The essential goal of data selection is assurance of appropriate data sort, source and instrument which enables experts to respond adequately to inquiries. For this construction we took both datasets as input from learning on the UCI machine learning repository.

3.1.2 Data preprocessing

Data cleansing is spinal cord of data science. A true data scientist always finds something out of the noisy data by the art of data cleansing. There are many techniques of doing data cleansing from which we opted to clean the noise out of raw data.

3.1.3 Training and testing of dataset

This is the part of implementation where we choose to do the partitioning of our data in to training and testing data. In our experiment, [70, 30] was the partition we used for our practical purposes because that's the standard partitioning ratio. [70, 30] means 70% of the data set is dedicated to training and 30% of data is dedicated to testing the algorithms if they predict the data being tested as accurately as possible, compared to their original number. This prediction is done on the basis of training data we have fed to the machine algorithms.

3.1.4 Apply K-fold Validation

K-Fold Validation is the most important step for verifying if the results given by the models are consistent and that is different sample of data is picked up randomly in the data set, accuracy won't be affected much. Procedure for K-Fold Validation was simple enough:

- **i.** Set the random seed value.
- **ii.** Set the partition as [70, 30], as kept in the first time practical.
- **iii.** Run the algorithms and generating the test result file.
- **iv.** Save the result file as csv file.
- **v.** Find the accuracy generated by different machine learning models in R.
- **vi.** Repeat steps 1 to 5 for 10 times.

3.1.5 Cross validation

In the simple words, testing sets are built just by part some unique dataset into more than one section. Yet, the assessments acquired for this situation have a tendency to mirror the specific way the data are isolated up. The arrangement is to utilize factual examining to get more precise estimations. This is known as cross-validation. The aim in cross-validation is to ensure that every example from the original dataset has the same chance of appearing in the training and testing set. The final output got from the cross validation to check robustness.

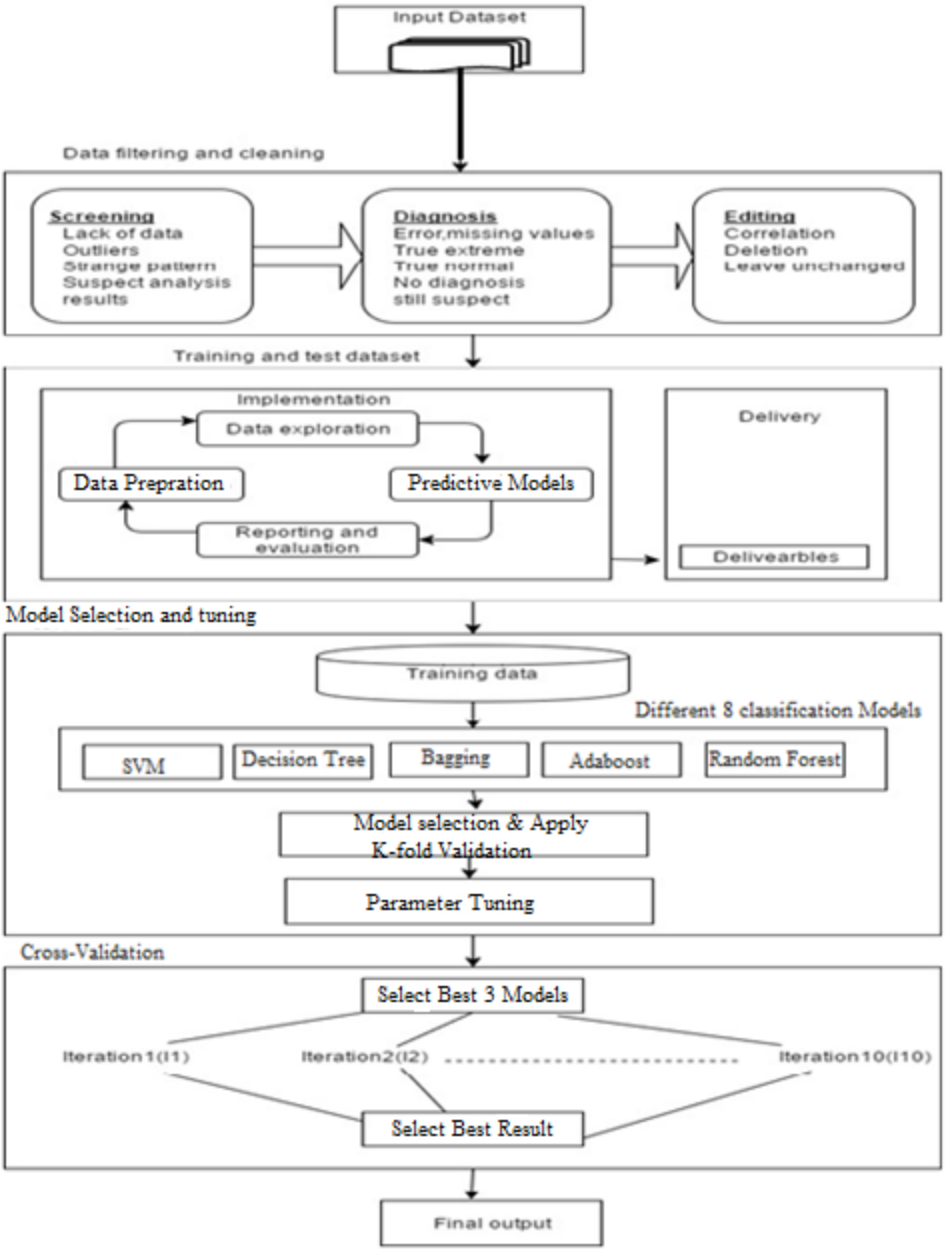

Fig 3.2: PaTM Framework

The concept of parameter tuning method is to analyze an evolutionary algorithm, and studying how the performance of the models varies when we are executing independent repetition of its runs. Along with K-cross validation is a method to calculate the accuracy of a system. For example, take the dataset, D, which is randomly divide into K equally exclusive subsets called folds of same size (D_1, D_2, …, D_k) and K classifiers are built. The i^{th} classifier is skilled on the addition of all value of j on D and checked on Di. The accuracy of the calculation is the overall number of the correct classification, which is

divided by the number of events occurring in the dataset. We have applied the seven classification models for predicting, testing and training, namely:

i) SVM
ii) Decision tree
iii) Logistic Regression
iv) Bagging
v) Stacking
vi) AdaBoost
vii) Random Forest

After this, output is utilized for training and testing by different classification algorithms separately and the result are shown in tabular and graphical form. The yield of all the seven algorithms is compared and analyzed and our result outperforms another state -of – the- art techniques.

Chapter 4: Design and Implementation of PaTM

This chapter describes the implementation done during the research, details of implementation of software, implementation of classification algorithms, and implementation of the proposed structure of analysis of parameter tuning and complete implementation snapshots.

4.1 Design of PaTM

The Software design is a procedure to renovate requirements of users into some appropriate form, which helps the programmer in coding and implementation of the software. Designing phase is the first step in SDLC (Software Design Life Cycle), which move the attention from problem domain to solution domain. It tries to identify how to accomplish the software requirements mentioned in Software Requirement Specification (SRS).

4.1.1 Architectural Design

The architectural design is the supreme essence of the system. This identifies the software as a system with many components interacting with each other. At the stage of architecture design, the designers get the idea of the proposed solution domain. The architecture of the proposed framework has shown in fig (3.2) chapter 3.

4.1.2 Activity diagram

The activity diagram in UML is another important diagram to describe the dynamic aspects of the system. Activity diagram is basically a flow chart representing the flow of activity from one activity to another. The activity can be described as the operation of the system. Control flow is drawn from one operation to another. This flow can be sequential, vegetarian or concurrent. Activity diagrams, to deal with all types of flow control, include various elements like fork, etc as shown in Figure (4.1). The overall details are discussed in previous chapter for each component of the activity diagram. The main aim of activity diagrams to captures the dynamic behavior of the system.

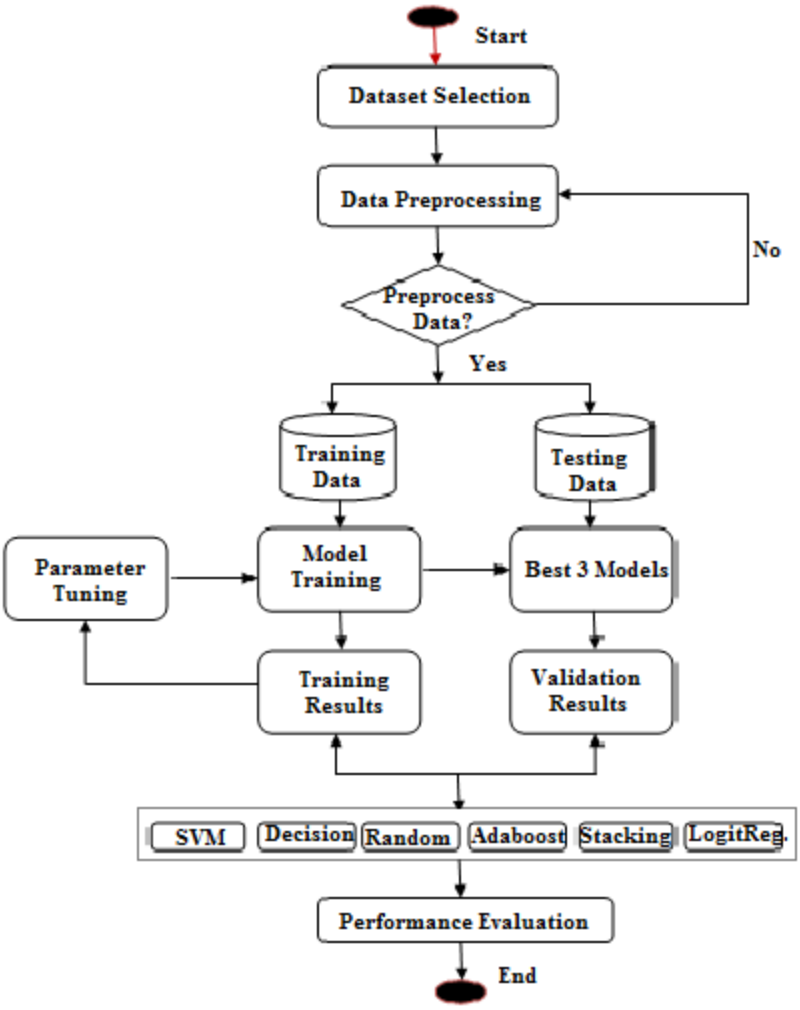

Fig 4.1: Activity Diagram for PaTM

4.2 Hardware and Software Requirements

Table 4.1: H/W and S/W Requirements

1.	RAM	4 GB
2.	System Type	64 bit/ operating system
3.	Hard Disk	500GB
4.	Operating System	Windows 10
5.	Programming Language	R (Rattle), Microsoft Excel
6.	Platform	R Studio

4.3 Module Overview

We have used the models tuning parameters module to build and test models using different combinations of parameters settings, in order to determine the optimum parameters for the given prediction task and data. In which the model's parameters as the actual values applied in the classification models, decision tree or other models generated by an algorithm. The models parameters are the settings and values you use when configuring and testing the model, with the aim of finding the best combinations. Model tune parameters support two methods for finding the optimum settings for a model:

i. **Integrated training and tuning:** Train a model using a parameter sweep. Most of the machine learning algorithms in which we can choose which parameters shall be changed during the training process, while other parameters remain fixed.

ii. **Cross validation with tuning:** In this method divide the data into some number of folds and then perform tests on the models to identify the best parameters for each subsection of the data. This method provides the best accuracy but can take longer to train. With both methods, you get an accuracy report describing the different models that were created and their parameters, trained a model that you can save for re-use.

4.3.1 Tuning Models Parameters

If we are not sure of the correct parameters for a given machine learning algorithm or task, we can conduct multiple parameters. We should also perform feature selection to determine the columns or variables that have the highest information value.

This section describes how to tuning the models parameter values:

i. Use the tune model parameters module to train a model while automatically finding the best parameters, using a parameter sweep.

ii. Perform cross-validation using the tune model parameters module.

4.3.2 Train Models with Parameter Tuning

In which we have train the different model parameters and add the module in experiment.

1. In the tuning models parameters, we have set options that define the number of parameters used and the number of iterations.

i. Select the option, specify the maximum number of runs that you want the model to execute. We have selected the parameter values over a system-defined range. This option is useful for those cases where we want to increase model performance.

2. The tune model parameters outputs provide a set of evaluation results, indicating the parameters that produced the best models, and the accuracy of all models.

4.3.3 Perform cross-validation

1. Perform the parameter tuning method in our classification models.
2. Select the datasets that we want to use for cross-validation. In which we do not change the parameter values, we need to divide the data into k-folds for cross-validation methods.
3. Choose the fold option and optionally specify some number of folds to divide the data into. By default, the parameter sweep performs 10-fold cross validation, with a random split. However, we can create any number of folds.
4. The parameter tuning models outputs bring a set of evaluation results, indicating the parameters that produced the best models, and their accuracy. The accuracy are calculated from the cross-validation pass, and may vary slightly depending on how many folds you selected.

Firstly, we have added the library rpart. After adding the library we have include the randomForest, ada, glm, lm and SVM packages, after that we performed read and write operation on the selected datasets.

4.4 Classification Models Implementation

We have predicted the accuracy of each and every classification model of Machine Learning used in this design.

4.4.1 SVM Implementation

We did all the implementation with the help of R language. For SVM implementation firstly we add various R packages and library like rpart, caret and fscaret. After adding rpart, caret and fscaret packages, we install SVM package and SVM library respectively. After that we performed read and write operation on imbalanced and balanced datasets and measure performance of the SVM in terms of confusion metrics.SVM can be implemented using following function which includes formula, trainDataset and method.

```
install.packages("e1071")
library(kernlab)
library(hmeasure)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- ksvm(formula, trainDataset, kernel="rbfdot", prob.model=TRUE)
```

After creating a model we measure the performance of created model using in terms of accuracy and ROC.

```
ConfusionMatrix <- misclassCounts(Predicted,Actual)$conf.matrix
# Evaluations Parameters
# AUC, ERR, Sen, Spec, Err,Pre,Recall, TPR, FPR, etc
EvaluationsParameters <- round(HMeasure(Actual,PredictedProb)$metrics,3)
#Accuracy
accuracy <- round(mean(Actual==Predicted) *100,2)
accuracy
```

4.4.2 Decision Tree Implementation

We did all the implementation with the help of R language. For SVM implementation firstly we add various R packages and library like rpart, caret and fscaret. After adding rpart, caret and fscaret packages, we install SVM package and SVM library respectively. After that we performed read and write operation on imbalanced and balanced datasets and measure performance of the SVM in terms of confusion metrics.SVM can be implemented using following function which includes formula, trainDataset and method.

```
install.packages("rpart")
library(rpart)
library(hmeasure)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model   <- rpart(formula, trainDataset, method="class", parms=list(split="information"),
                 control=rpart.control(usesurrogate=0, maxsurrogate=0))
```

After creating a model we measure the performance of created model using in terms of accuracy and ROC.

```
# Evaluations Parameters
# AUC, ERR, Sen, Spec, Pre,Recall, TPR, FPR, etc
EvaluationsParameters <- round(HMeasure(Actual,PredictedProb)$metrics,3)
EvaluationsParameters
# Accuracy
accuracy <- round(mean(Actual==Predicted) *100,2)
accuracy
```

4.4.3 AdaBoost Implementation

We did all the implementation with the help of R language. For AdaBoost implementation firstly we add various R packages and library like rpart, caret and fscaret. After adding rpart, caret and fscaret packages, we install ada package and ada library respectively. After that we performed read and write operation on imbalanced and balanced datasets and measure performance of the adaboost in terms of confusion metrics. AdaBoost can be implemented using following function which includes formula, trainDataset.

```
install.packages("ada")
library(ada)
library(hmeasure)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model <- ksvm(formula, trainDataset, kernel="rbfdot", prob.model=TRUE)
model <- ada(formula, trainDataset)
                     control=rpart::rpart.control(maxdepth=30,
                                                  cp=0.010000,
                                                  minsplit=20,
                                                  xval=10)
                                                  iter=50
```

After creating a model we measure the performance of created model using in terms of accuracy and ROC.

```
ConfusionMatrix <- misclassCounts(Predicted,Actual)$conf.matrix
# Evaluations Parameters
# AUC, ERR, Sen, Spec, Pre,Recall, TPR, FPR, etc
EvaluationsParameters <- round(HMeasure(Actual,PredictedProb)$metrics,3)
EvaluationsParameters
# Accuracy
accuracy <- round(mean(Actual==Predicted) *100,2)
accuracy
```

4.4.4 Random Forest Implementation

We did all the implementation with the help of R language. For Random Forest implementation firstly we add various R packages and library like rpart, caret and fscaret. After adding rpart, caret and fscaret packages, we install randomForest package. After that we performed read and write operation on imbalanced and balanced datasets and measure performance of the random forest in terms of confusion metrics. Random forest can be implemented using following function which includes formula, train dataset and method.

```
install.packages("randomforest")
library(randomForest)
library(hmeasure)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model   <- randomForest(formula, trainDataset, ntree=500,mtry=2)
```

After creating a model we measure the performance of created model using in terms of accuracy and ROC.

```
ConfusionMatrix <- misclassCounts(Predicted,Actual)$conf.matrix
# Evaluations Parameters
# AUC, ERR, Sen, Spec, Pre,Recall, TPR, FPR, etc
EvaluationsParameters <- round(HMeasure(Actual,PredictedProb)$metrics,3)
EvaluationsParameters
# Accuracy
accuracy <- round(mean(Actual==Predicted) *100,2)
accuracy
```

Chapter 5: Experimental Results

The project code has been done in R language and weka tool and the pretension have done on Windows 10, 64 bit machine. Improve the performance of classification models in terms of accuracy and ROC and reduce the error rate of the classification models. To select the different dataset and applying 7 different classifier models; the datasets belongs to the classification data like binary dataset and multiclass datasets so we have applied some Machine Learning model based on the classification. The summary of datasets display in the following Table (5.1)

Table 5.1: Dataset Details

Sr. No	**Dataset Names**	**No. of Instances**	**No. of Attribute**	**No. of Class**
1.	Diabetes	768	9	2
2.	Vote	435	17	2
3.	Breast Cancer	286	10	2
4.	Credit Cards	1000	21	4
5.	Nursery Dataset	12960	9	5
6.	Vehicles	846	19	4

The methodology is divided into seven different levels and each level is described as:

5.1 Methodology

Level 1: In first level, selected the different datasets based on classification. In which different types of classification datasets like binary class datasets and multiclass datasets with no. of instance, no. of attributes, and no of class.

Level 2: In the second phase, preprocessing of the dataset like removal of duplicate, noise and missing value entries on different datasets. There are different tasks able in data preprocessing for e.g. cleaning, transformation, reduction, and data integration. These all tasks will be performed carried in second phase.

Level 3: In third level, applied 8 classification machine learning algorithms like SVM, Decision tree, Logistic regression, Bagging, Stacking, Adaboost, Random forest can be used to train and test the dataset Studio.

Level 4: In this level, trained the classifier using the training set, tune the performance of the classifier using the validation set and then finally test the performance of our models.

Level 5: In this level, using different K-fold cross-validation would be used to measure the robustness of the best predictive method.

Level 6: This is the evaluation phase where the different parameter values of different classifier have been changed and used these terms like accuracy, error rate, ROC to evaluate the performance of the models.

Level 7: This is the last step of our experiments in which finally, we would get the results in terms of accuracy, error rate, and ROC.

5.2 Results

An experimental result of proposed framework is presented in following Tables and Figures. The table contains outcome of 7 different classification models along with the value of Accuracy and ROC. The suggested framework is applied on different classification datasets. Compared the all these algorithms on the basis of their accuracy and ROC. The experimental results are shown in following Table 5.2 and Figure 5.3.

5.2.1 Performance evaluation using Accuracy

As we have already explained that the classification algorithms applied on the different UCI repository dataset. These algorithms are as follows: SVM, Decision Tree, Logistic regression, Bagging, stacking, AdaBoost, and Random forest. To use the 10-folds validation method and compare all these models performance on the basis of their accuracy.

Table 5.2: Performance evaluation using Accuracy

Sr.no	Datasets	Best 3 Models	Default Accuracy	Improved Accuracy	Best Accuracy
1.	Diabetes	Random Forest SVM Bagging	72.91 77.34 76.30	76.82 77.96 77.34	77.96
2.	Vote	SVM Stacking Random Forest	96.09 98.32 94.09	97.08 96.64 96.32	96.64
3.	Brest Cancer	Simple Logistic Logit Boost Stacking	75.17 72.37 75.52	76.77 75.52 75.82	76.77

4.	Credit Card	AdaBoost Simple Logistic Stacking	76.4 75.9 75.2	78.4 77.4 77.4	78.4
5.	Nursery Dataset	Bagging Random Forest Decision Tree	97.33 98.06 97.05	98.53 99.09 99.53	99.53
6.	Vehicles	Simple Logistic Stacking AdaBoost	77.18 76.78 76.00	77.34 79.97 77.4	79.97

Here, the experimental results of the Table 5.2 are shown in the graphical form they are as:

Now to train the Diabetes dataset and applied the validation along with parameter tuning. After testing the dataset process, we have used different K-fold cross-validation method to measure the robustness of the best predictive method. Now again we have applied the classification algorithms on Diabetes dataset. After applying the parameter tuning method on classification algorithms the Diabetes dataset results are improved as default performance results of Diabetes dataset. The experimental results are presented in Figure 5.1.

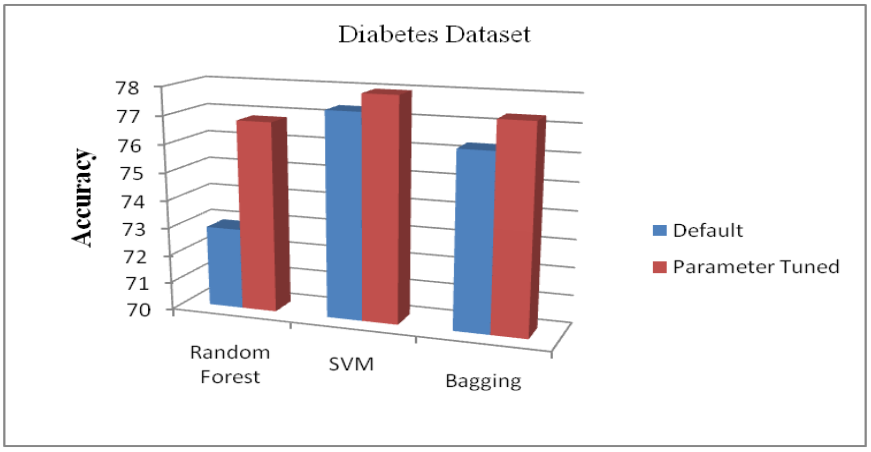

Fig 5.1: Diabetes Dataset

In which the different classification algorithms have been applied on Vote prediction dataset after tuning the parameter values. The experimental results prove that the performance of the applied models has been improved from the default result of the Vote prediction dataset. The result presented in Figure 5.2.

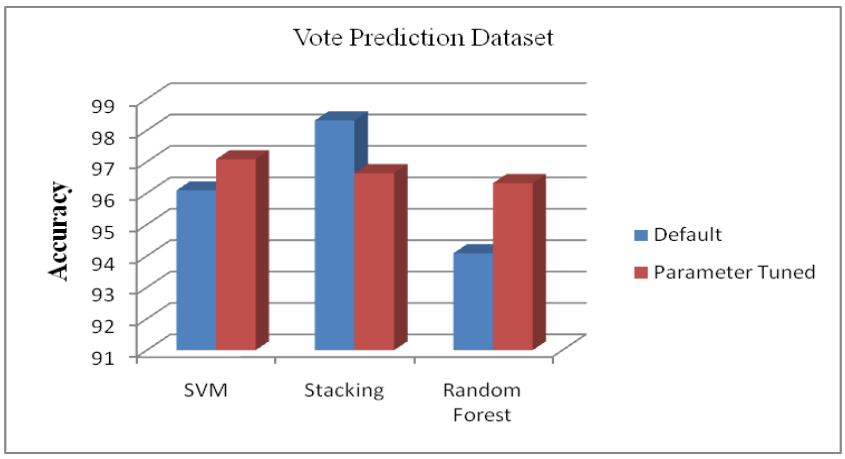

Fig 5.2: Vote Prediction Dataset

After tuning the parameter values applied the classification models on Breast Cancer dataset, the experimental results are presented in Figure 5.3

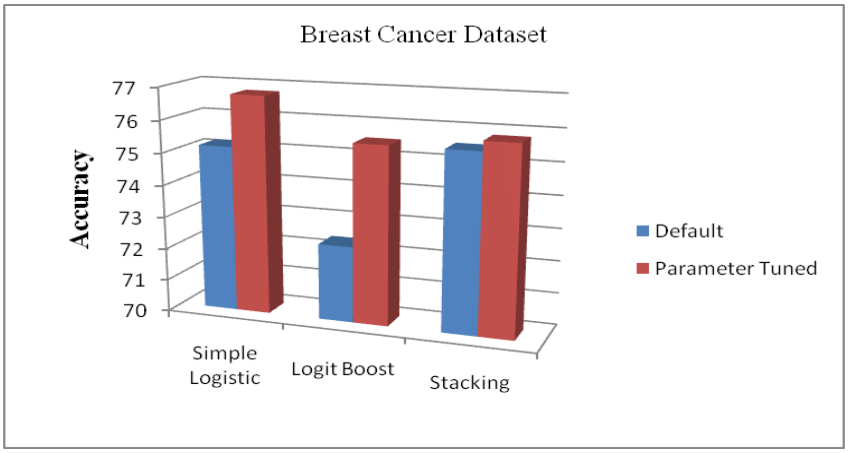

Fig 5.3: Breast Cancer Dataset

Various classification algorithms have been applied on Credit cards dataset. After tuning the parameters and applying classification algorithms on Credit card dataset, the experimental results are presented in Figure 5.4.

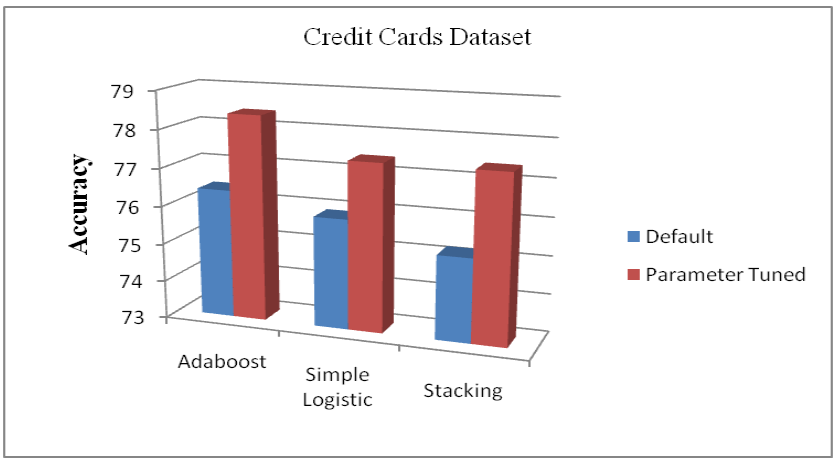

Fig 5.4: Credit Cards dataset

We have applied the classification algorithms on Nursery dataset. After changing the parameter values and applying again the classification algorithms on Nursery dataset, the experimental result prove that the performance of the models has been improve after tuning the values. The results presented in Figure 5.5.

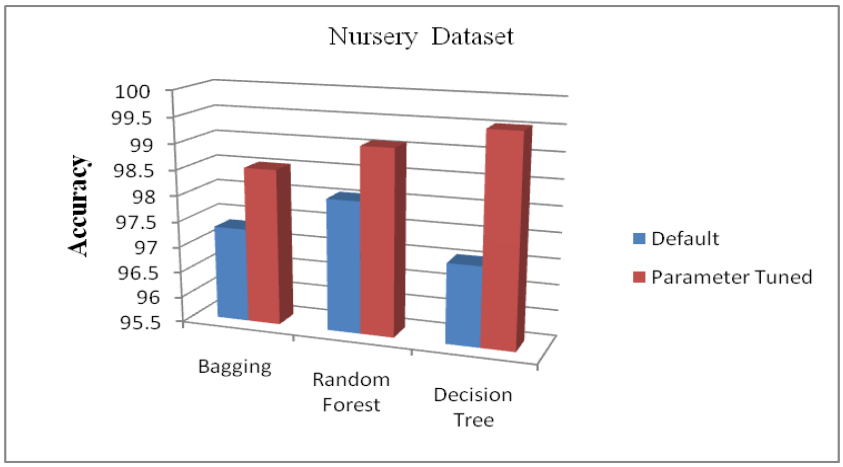

Fig 5.5: Nursery Dataset

Using the various classification models and applied these models on the Vehicle dataset after tuning the parameter values. The performances of the results are improved from the default classification models. The results presented in Figure 5.6.

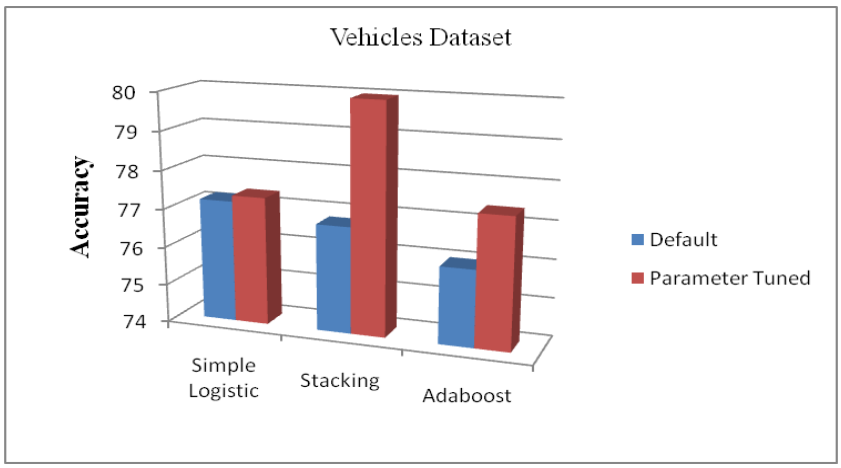

Fig 5.6: Vehicles Dataset

5.1.2 Performance evaluation using ROC

Here, below table contains the same process used in the accuracy Table 5.2. In which the outcome of 7 different Machine Learning classification models along with the term of ROC. We have been used the 10-folds validation method and compared all these algorithms on the basis of their accuracy.

Table 5.3: Models Evaluation Results in Term of ROC

Sr.no	Datasets	Best 3 Models	ROC	Improved ROC
1.	Diabetes	Random Forest SVM Bagging	0.81 0.72 0.79	0.82 0.71 0.78
2.	Vote	SVM Stacking Random Forest	0.96 0.97 0.95	0.97 0.98 0.97
3.	Breast Cancer	Simple Logistic Logit Boost Stacking	0.67 0.69 0.58	0.69 0.72 0.66
4.	Credit Card	Adaboost Simple Logistic Stacking	0.79 0.79 0.72	0.81 0.80 0.75
5.	Nursery Dataset	Bagging Random Forest Decision Tree	0.98 0.99 0.97	0.99 1.0 0.97
6.	Vehicles	Simple Logistic Stacking Adaboost	0.93 0.94 0.93	0.95 0.96 0.95

Here, the experimental results of the Table 5.2 are shown in the graphical form they are as:

At the same process, in the above method in which we have used different K-Fold cross-validation method to measure the robustness of the best predictive method. Now again we have applied the classification algorithms on Diabetes dataset. After applying the parameter tuning method on classification algorithms the Diabetes dataset results are improved as default performance results of Diabetes dataset in term of ROC. The experimental results are presented in Figure 5.7.

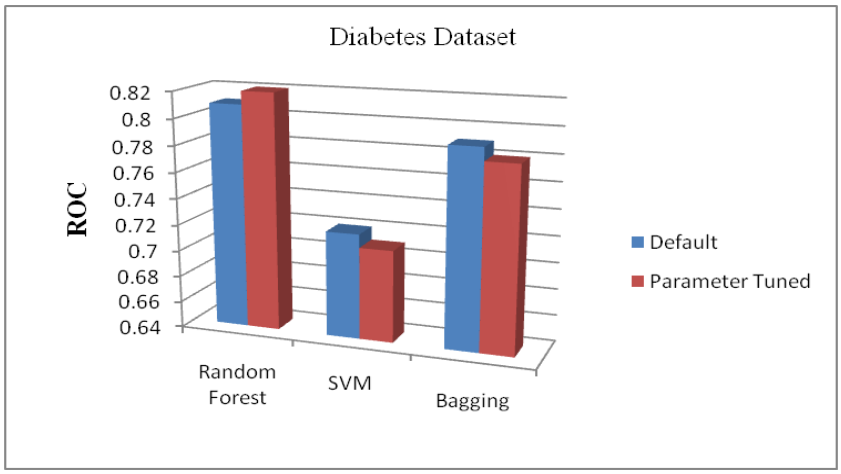

Fig 5.7: Diabetes Dataset

Numerous classification models have been applied on Vote prediction dataset. Then perform the tuning method and applying classification algorithms on Vote prediction dataset, the experimental results are presented in Figure 5.8.

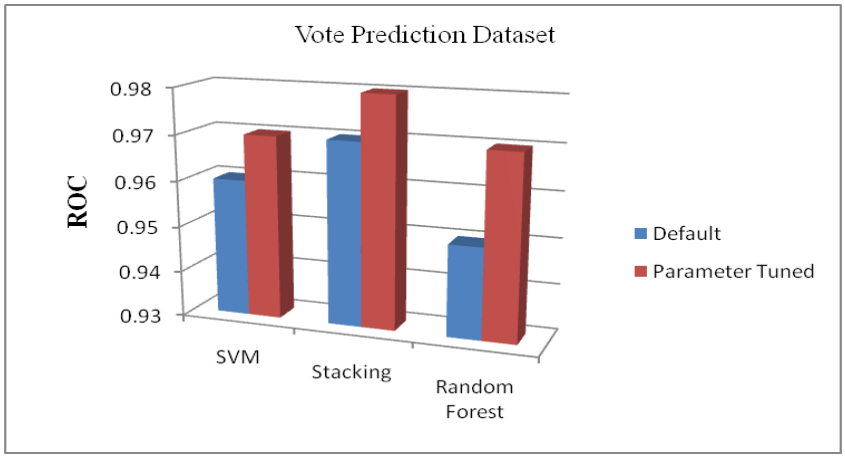

Fig 5.8: Vote Prediction Dataset

Various classification algorithms have been applied on Breast cancer dataset. After tuning the parameters and applying classification algorithms on Breast cancer card dataset, the experimental results are presented in Figure 5.9.

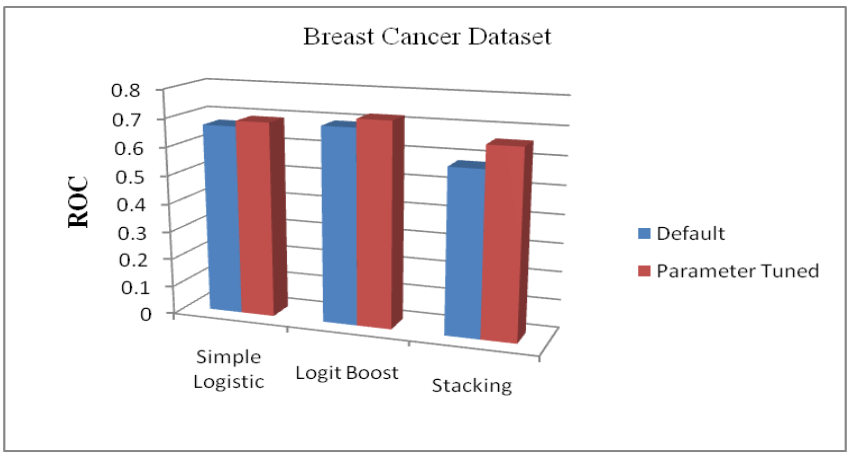

Fig 5.9: Breast Cancer dataset

We have applied the classification algorithms on Credit cards dataset. After tuning the parameters and applying classification algorithms on Credit card dataset, the experimental results are presented in Figure 5.10.

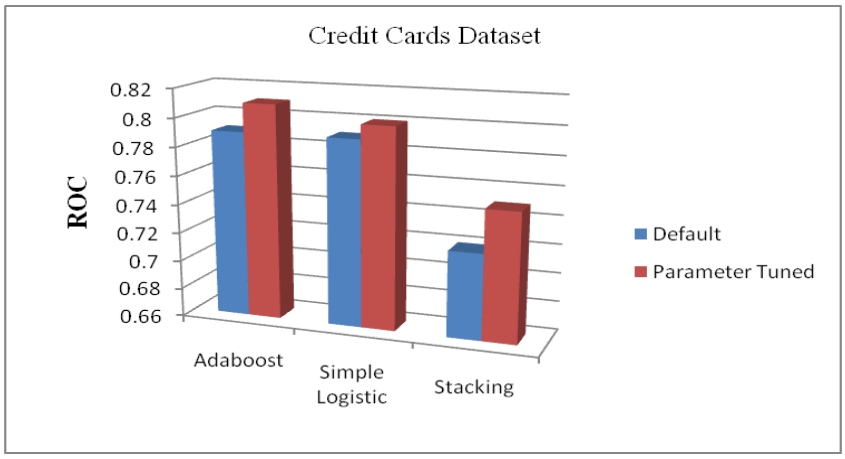

Fig 5.10: Credit Cards Dataset

In which different classification algorithms have been applied on Nursery dataset. Then performing tuning method and again applied classification algorithms on Nursery dataset, the experimental results are presented in Figure 5.4.

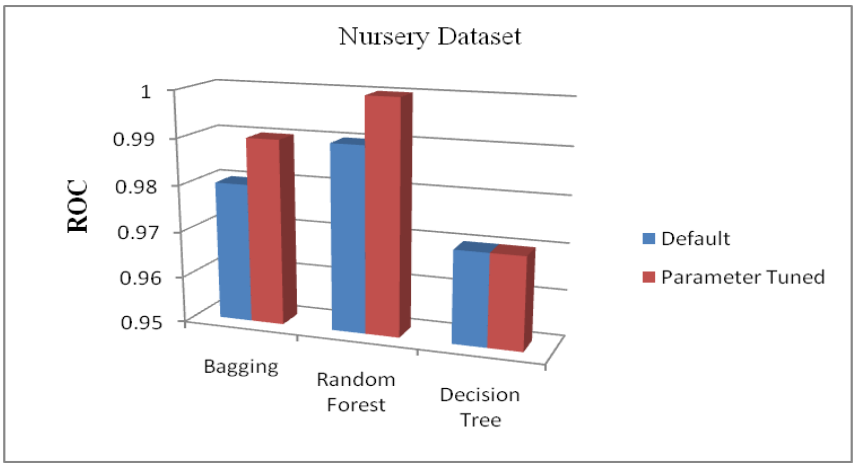

Fig 5.11: Nursery Dataset

We have applied the classification algorithms on Vehicles dataset. After tuning the parameters and applying classification algorithms on Vehicles dataset, the experimental results are presented in Figure 5.12.

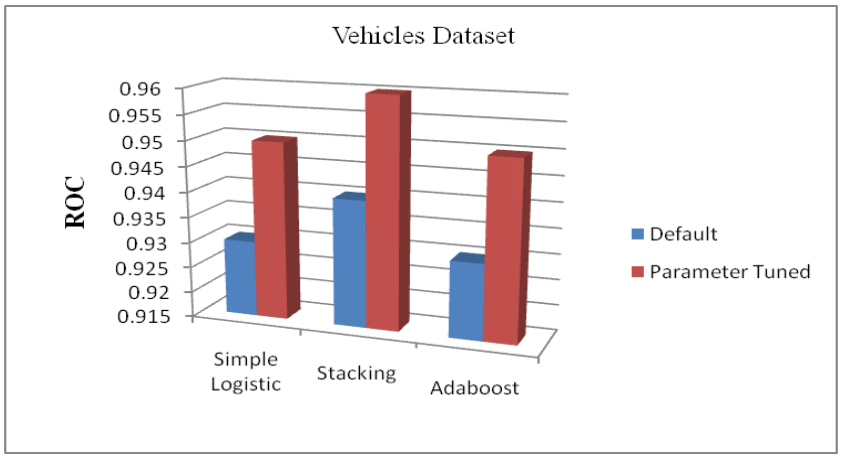

Fig 5.12: Vehicles Dataset

Chapter 6: Conclusion and Future Scope

6.1 Conclusion

Searching a good configuration for the parameters of machine learning algorithms requires trials, intuition as well as a specialized knowledge. Tuning of parameters is seen as an optimization problem, whose purpose is to optimize the estimated performance of the motivated model (e.g., accuracy) by algorithm. To construct an evolutionary algorithm through the parameter tuning method by selecting the best parameter values that optimizes the performance of the classification models. Learning the classification models to examine how the performance of the models depends on its parameter values. The performance of the models has been improvised using proposed parameter tuning framework. Various classification models like SVM, Random Forest, and Decision Tree etc. has been tested to improve the performance of the models and also to change the important parameter values of the above models. Cross10-fold validation method is also used to check the performance of the models by applying number of iterations on the particular dataset. The classifier's algorithms are given the best results in terms of accuracy after applying the independent repetitions of the single dataset. After that, we have applied different classifiers on other different datasets with the help of various R packages and libraries by selcting important parameter values. Further, we have applied classification algorithms after changing the parameter values and check the accuracy and ROC improvements of the models. It can be concluded that classification of data can be improved significantly to key out the rare events from the datasets by applying the parameter tuning method. Now finally, it is concluded that proposed framework has shown the better performance results.

6.2 Future work

In future work, we have the intention to calibrate different approaches to extract the important parameter and expand the number of data sets. We also plan to explore other methodologies on the Meta level, including data balancing and meta-feature selection process to find the most relevant meta-features. Moreover, we should include a

significance test to define the meta-target, and include experiments with other classification algorithms and Deep Learning algorithms, which have a larger number of sensitive hyper-parameters.

References

[1] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1. No. 1. Cambridge: MIT press, 1998.

[2] Rasmussen, Carl Edward. "Gaussian processes in machine learning." *Advanced lectures on machine learning*. Springer Berlin Heidelberg, 2004. 63-71.

[3] "Supervised and Unsupervised Machine Learning Algorithms." *Machine Learning Mastery*. N.p., 21 Sept. 2016. Web. 26 June 2017.

[4] Www.facebook.com/sandeep.dayananda. "Spark MLlib | Machine Learning In Apache Spark | Spark Tutorial | Edureka." *Edureka Blog*. N.p., 10 May 2017. Web. 26 June 2017.

[5] "Brief Review of Regression-Based and Machine Learning ..." N.p., n.d. Web. 26 June 2017.

[6] "Getting started with Classification." *GeeksforGeeks*. N.p., 15 June 2016. Web. 26 June 2017.

[7] R. Ranawana, V. Palade, "Optimized precision: a new measure for classifier performance evaluation", IEEE Congress on Computational Intelligence, Canada, pp. 2254–2261, 2006.

[8] M. K. Taghi, V. H. Hulse, N. Amri, "Comparing boosting and bagging techniques with noisy and imbalanced data", IEEE Transactions Systems, Man, Cybernetics A Systems Humans, pp. 552-568, 2011.

[9] K. Priyanka, N. Abhigyan, C. Radha, "Identification of human drug targets using machine-learning algorithms", ELSEVIER Computers in Biology and Medicine 56, pp. 175–181, 2015.

[10] "An Introduction to Feature Selection." *Machine Learning Mastery*. N.p., 30 Oct. 2016. Web. 26 June 2017.

[11] Eiben, Agoston E., and Selmar K. Smit. "Parameter tuning for configuring and analyzing evolutionary algorithms." *Swarm and Evolutionary Computation* 1.1 (2011): 19-31.

[12] Mantovani, Rafael G., et al. "To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning." *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015.

[13] Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. Vol. 1. No. 1. Cambridge: MIT press, 1998

[14] Villa, Alberto, et al. "Gradient optimization for multiple kernel's parameters in support vector machines classification." *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*. Vol. 4. IEEE, 2008.

[15] Sujatha, M., S. Prabhakar, and Dr G. Lavanya Devi. "A Survey of Classification Techniques in Data Mining." *International Journal of Innovations in Engineering and Technology (IJIET)* 2.4 (2013).

[16] T. M. Mitchell, "The discipline of Machine Learning", Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Dept., 2006.

[17] Shalev-Shwartz, Shai, and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] Austin, Peter C., et al. "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes." *Journal of clinical epidemiology* 66.4 (2013): 398-407.

[19] Langley, Pat, and Herbert A. Simon. "Applications of machine learning and rule induction." *Communications of the ACM* 38.11 (1995): 54-64.

[20] Rasmussen, Carl Edward. "Gaussian processes in machine learning." *Advanced lectures on machine learning*. Springer Berlin Heidelberg, 2004. 63-71.

[21] Woon, Wei Lee, Ayman El-Hag, and Mustafa Harbaji. "Machine learning techniques for robust classification of partial discharges in oil–paper insulation systems." *IET Science, Measurement & Technology* 10.3 (2016): 221-227.

[22] Najafabadi, Maryam M., et al. "Deep learning applications and challenges in big data analytics." *Journal of Big Data* 2.1 (2015): 1.

[23] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24

.

[24] Chitra, R., and V. Seenivasagam. "Heart disease prediction system using supervised learning classifier." *Bonfring International Journal of Software Engineering and Soft Computing* 3.1 (2013): 1.

[25] Kotsiantis, Sotiris B., Ioannis D. Zaharakis, and Panayiotis E. Pintelas. "Machine learning: a review of classification and combining techniques." *Artificial Intelligence Review* 26.3 (2006): 159-190.

[26] Yang, Tao, Dongmei Fu, and Lian Hao. "Supervised laplacian graph multiple kernel classification." *Society of Instrument and Control Engineers of Japan (SICE), 2016 55th Annual Conference of the*. IEEE, 2016.

[27] Figueroa, Alejandro, and John Atkinson. "Ensembling Classifiers for Detecting User Intentions behind Web Queries." *IEEE Internet Computing* 20.2 (2016): 8-16.

[28] Sujatha, M., S. Prabhakar, and G. A. Devi. "A Survey of Classification Techniques in Data Mining." *International Journal of Innovations in Engineering and Technology (IJIET)* 2.4 (2013).

[29] Huang, Gao, et al. "Semi-supervised and unsupervised extreme learning machines." *IEEE Transactions on Cybernetics* 44.12 (2014): 2405-2417.

[30] Pang, Guansong, et al. "Unsupervised Feature Selection for Outlier Detection by Modelling Hierarchical Value-Feature Couplings." *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016.

[31] Ferreira, Pedro M. "Unsupervised entropy-based selection of data sets for improved model fitting." *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016.

[32] Zhang, Shaohua, Hua Yang, and Zhouping Yin. "Performance evaluation of typical unsupervised feature learning algorithms for visual object recognition." *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, 2014.

[33] Asif, Muhammad Tayyab, et al. "Unsupervised learning based performance analysis of n-support vector regression for speed prediction of a large road network." *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*. IEEE, 2012.

[34] Li, Xuchun, Lei Wang, and Eric Sung. "A study of AdaBoost with SVM based weak learners." *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. Vol. 1. IEEE, 2005.

[35] Assareh, Amin, L. Gwenn Volkert, and Jing Li. "Feature selections using AdaBoost: Application in gene-gene interaction detection." *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*. IEEE, 2012.

[36] Hashi, Emrana Kabir, Md Shahid Uz Zaman, and Md Rokibul Hasan. "An expert clinical decision support system to predict disease using classification techniques." *Electrical, Computer and Communication Engineering (ECCE), International Conference on*. IEEE, 2017.

[37] Joshi, Nikita, and Shweta Srivastava. "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)." (2014)

[38] Mantovani, Rafael G., et al. "Hyper-parameter Tuning of a Decision Tree Induction Algorithm." *Brazilian Conference on Intelligent Systems (BRACIS 2016)*. 2016..

[39] Wang, Guolu, Jungang Xu, and Ben He. "A Novel Method for Tuning Configuration Parameters of Spark Based on Machine Learning." *High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on*. IEEE, 2016.

[40] Mantovani, Rafael G., et al. "To tune or not to tune: recommending when to adjust SVM hyper-parameters via Meta-learning." *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015.

[41] Ding, Sheng, and Shunxin Li. "PSO parameters optimization based support vector machines for hyperspectral classification." *Information Science and Engineering (ICISE), 2009 1st International Conference on*. IEEE, 2009.

[42] Novakovic, J., and A. Veljovic. "C-support vector classification: Selection of kernel and parameters in medical diagnosis." *Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on*. IEEE, 2011.

[43] Eiben, Agoston E., and Selmar K. Smit. "Parameter tuning for configuring and analyzing evolutionary algorithms." *Swarm and Evolutionary Computation* 1.1 (2011): 19-31.

[44] Gao, Hengzhen, Mrinal K. Mandal, and Jianwei Wan. "Classification of hyperspectral image with feature selection and parameter estimation." Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on. Vol. 1. IEEE, 2010.

[45] Chiu, Chien-Yuan, Brijesh Verma, and Michael Li. "Impact of variability in data on accuracy and diversity of neural network based ensemble classifiers." Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013.

[46] Molina, M. M., et al. "Meta-Learning Approach for Automatic Parameter Tuning: A Case Study with Educational Datasets." International Educational Data Mining Society (2012).

[47] Fei, Ye, and Han Min. "Simultaneous feature with support vector selection and parameters optimization using GA-based SVM solve the binary classification." Computer Communication and the Internet (ICCCI), 2016 IEEE International Conference on. IEEE, 2016.

[48] Sherin, B. M., and M. H. Supriya. "Selection and parameter optimization of SVM kernel function for underwater target classification." Underwater Technology (UT), 2015 IEEE. IEEE, 2015.

List of Publication

[1] Lata Dubey and Seema Bawa, **"*Parameter tuning Method Analytics for Different Datasets using Classification Models"***, in *IJRSR International Journal of Recent Scientific Research, Impact and Indexing Journal-2017-IJRSR* [Communicated].

Plagiarism Report

Lata-801532028

ORIGINALITY REPORT

%4	%3	%2	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	gohcy.com Internet Source	%1
2	Shaohua Zhang, , Hua Yang, and Zhouping Yin. "Performance evaluation of typical unsupervised feature learning algorithms for visual object recognition", Proceeding of the 11th World Congress on Intelligent Control and Automation, 2014. Publication	%1
3	users.cs.cf.ac.uk Internet Source	%1
4	file.scirp.org Internet Source	<%1
5	Mantovani, Rafael G., Andre L. D. Rossi, Joaquin Vanschoren, Bernd Bischl, and Andre C. P. L. F. Carvalho. "To tune or not to tune: Recommending when to adjust SVM hyper-parameters via meta-learning", 2015 International Joint Conference on Neural Networks (IJCNN), 2015.	<%1