Diverse Ensemble Framework (DEF) For Predictive Analytics

A Thesis

submitted in partial fulfilment of the requirements for the award of the degree of

Master of Engineering

in

Software Engineering

by

Ashish Gour

(Roll No: 801531005)

Under the supervision of

Dr. Seema Bawa

(Professor)

Computer science and Engineering Department

THAPAR UNIVERSITY

PATIALA-147004, PUNJAB, INDIA

June 2017

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Diverse Ensemble Framework for Predictive Analytics*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted 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 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.

(Ashish Gour)

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

22/06/2017

(Dr. Seema Bawa)
Professor, CSED

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. She has 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 her. I also thank my supervisor for her time, patience, discussions and valuable comments. Her enthusiasm 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. Rupali Bhardwaj**, 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.

Ashish Gour

(801531005)

Abstract

Machine Learning technique has numerous benefits that include high flexibility and power, lack of parametric assumption etc. Building an effective Machine Learning ensemble model, by using different categories models to perform ensembling than the concept of diversity in ensemble has occurs. To perform ensembling on the datasets the technique like bagging, boosting, and voting are used. In this thesis we take the two datasets of regression data and initially execute both the datasets on eight different machine learning models (RF, NN, LM, Cubist, Enet, LR with SS, PCR and ICR), after performing ensemble techniques on the following machine learning model we got the batter result as compare to individual results.

In classification problems, observations fall into pre-assigned groups. Examples include identifying customers who would buy a product, and detecting whether a credit card expense is made by a customer. A popular approach to tackle these problems is using a collection of models that combines the collective knowledge of them. It has been shown that employing multiple models outperforms a single model. A common approach has been to use the same collection for all observations, which is also known as the static approach. Recently, there have been more attempts in using a different collection that is more specialized for each observation, depending on the features of observations.

In the ensemble result seen that the ensemble of two weak models (with minimum accuracy) are gives the best results (with high accuracy) in the both the case of datasets. To make the work more efficient we proposed a framework known as Diversity framework (DEF). The proposed framework is also predicting the odor of the chemicals so it is known as olfaction prediction. The framework DEF is developed using R language and various R packages. The performance of DEF is evaluated and results show that DE framework out performs than other existing techniques like Bagging, Boosting etc.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This chapter discussed about Ensemble, Diversity, and Machine Learning techniques to deal with diversity in ensembles.

1.1 Machine Learning Overview

The goals of Machine Learning are to learn complicated patterns and to make intelligent decisions based on input data without human intervention. “The ability of a program to learn from experience is to modify its execution on newly acquired information”. If computer science is the standardized characterization of estimation that what we can perform conveniently then, is Machine Learning? To deal with a problem in a computer, one first plan an appropriately competent algorithm that deals with the problem and then designs and implements that algorithm in software or hardware. We cannot solve the problem without implement and design an algorithm for that problem. When we are unable to solve a problem manually then Machine Learning extend what can we do with a computer, and how we play with the programmed algorithm.

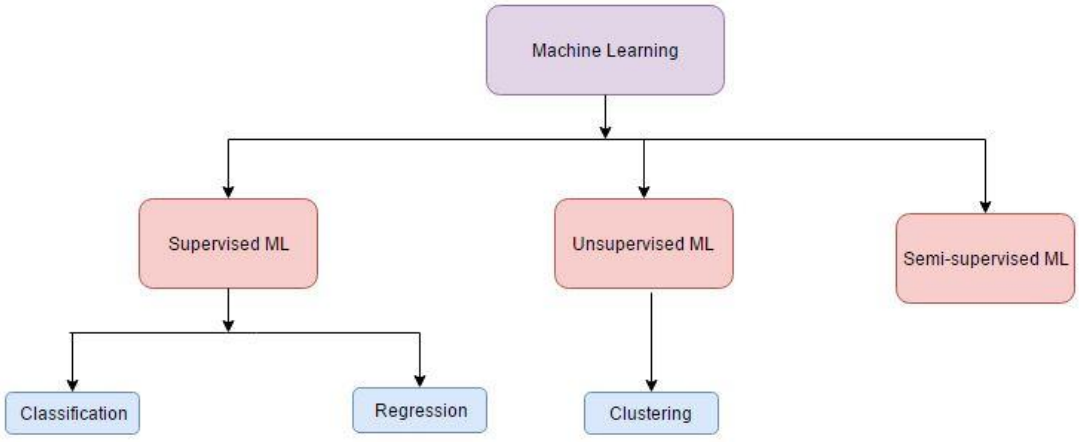

Figure 1.1 Types of Machine Learning

Examining Machine Learning helps us to comprehend what we can process for all intents and purposes, we should learn about calculation keeping in mind the end goal to

educate our common comprehension with respect to learning. Machine Learning as a logical train inspects the computational premise of learning; consequently, it is basic regardless of the possibility that we are just keen on how people and creatures learn. These two focal inspirations bolster each other. Attempting to take care of issues utilizing computational models of learning reveals insight into our comprehension of the cerebrum, and, by a similar token, what we find out about the mind can fill in as motivation for outlining learning machines. Types of Machine Learning are shown in Figure (1.1)

1.1.1 Types of Machine Learning

A. **Supervised learning**: A supervised learning algorithm breaks down the preparation information and produces a derived capacity, which can be 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. For example, you have input thing (x) and a yield thing (Y) and you utilize a calculation to take in the processing capacity called yield with the assistance of info.

$$Y = f(X) \qquad \text{……………………………………………..…eq. (1)}$$

The point is to inexact the registering capacity so fine that after you utilize new info thing (x) that you can foresee yield factors (Y) for that information. Supervised learning problem be able to group into classification and regression problem.

i. **Classification**: The goal of the classification algorithm is to forecast the target set yes or no, for predicting two target value or class we use binary classification, i.e. to predict student profile status fail or pass. When we have predicted for more than two target data class we use the multiple classifications, i.e. considering all the student details of the student, to estimate which students will earn more points.

ii. **Regression**: The goal of regression algorithm is to predict continuous or discrete values. Once in a while, the foreseeing quality can be utilized to locate the straight connection between the attributes. Basic regression algorithm such as linear, polynomial, etc is used in Machine Learning problems. Some famous regression algorithm of supervised learning are follows

a) **Linear Regression**: It is utilized to gauge genuine esteems (cost of houses, the quantity of calls, aggregate deals and so on.) in view of a persistent variable(s). Here, we set up a connection amongst autonomous and ward factors by fitting the best line. This best fit line is known as relapse line and spoken to by a direct condition.

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

b) **Decision Tree:** Decision tree learning utilizes a decision tree as a predictive model perceptions 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.

B. **Unsupervised learning**: The ambition of unsupervised learning is in the direction of model the underlying structure or distribution in the data in order to learn more about the data. Unsupervised learning is a type of Machine Learning algorithm that draws references from a dataset with input data without labeled responses. It is distinguished from supervised learning (and reinforcement learning) in that the learner is given only unlabeled examples. Unsupervised learning problems can be further grouped into clustering and association problems.

i. **Clustering**:

The clustering issue is a place you need to find the innate groupings in the information, for example, gathering clients by obtaining conduct. Clustering is the mission of an arrangement of perceptions into subsets (called bunches) so that perceptions in a similar bunch are comparable in some sense. Clustering is a strategy for unsupervised learning and a typical procedure for factual information investigation utilized as a part of many fields.

ii. **Association**:

An association rule learning issue is the place you need to find decides that portray expensive segments of your information, for example, individuals that

purchase X additionally tend to purchase Y. Some prominent cases of unsupervised learning calculations are:

a. K-means used for clustering evils.
b. Apriori algorithm for the association rule mining.

1.2 Ensemble Methods

Ensemble method is the combination of various classification algorithms to enhance prediction accuracy and the ability of generalization. Most prominent ensemble-based strategies are packing and boosting. In boosting, order algorithm utilizes past one and furthermore concentrates on its mistakes, while packing trains every characterization algorithm by a subset of the preparation set. The different ensemble method discussed in Table 1.1 Ensemble techniques (Jain *et al.* 2000).

Table 1.1 Ensemble techniques

S.no	Techniques	Architecture	Trainable	Adaptive	Information level
1.	Voting	Parallel	No	No	Abstract
2.	Bagging	Parallel	Yes	No	Confidence
3.	Boosting	Parallel hierarchical	Yes	No	Abstract
4.	Random subspace	Parallel	Yes	No	Confidence
5.	Borda Count	Parallel	Yes	No	Rank
6.	Stacking	Gated Parallel	Yes	Yes	Confidence

The combination of Boosting, Bagging, and stacking is called "Meta-Algorithm approach" is used to decrease the variance or improve the predictive force of the ensembles. Some of the above techniques are explained below.

1.2.1 Bagging

Bagging (Bootstrap aggregation) [32] 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) [33]. 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 [28]. 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 dress with REPTree is used to implement intrusion detection system [38].

1.2.2 Boosting

Boosting [34] has a whole family of equal family members, such as winning, utilizing voting in favor of coalitions to join the forecasts of a base model learned via a solitary learning algorithm. The contrast in among two methodologies with the intention of the built-in base models are dropped on the occasion of completing, while we try to model the supplementary model by learning further models, keeping in mind the mistakes of the previous model. With the same learning examples, learning the a respectable starting point display on the whole showing set begins the procedure. For the following base model, we need to get an exact gauge of the illustrations which have not been appropriately anticipated by past base models. Accordingly, we increment the heaviness of these cases (or shed pounds of exact prescient illustrations) and take in another base model. When some stop criteria are satisfied, we stop learning new base models.

1.2.3 Stacking

Stacking [37] or stacked speculation is the strategy for consolidating odd base machine models, i. e., models were found out by various learning techniques, for example, the nearest neighboring strategy, choice tree, unconstrained bayas and so forth. Base models are not consolidated with an unequivocal arrangement like voting, rather an extra model known as Meta or stage model is found out and utilized for the blend of the base or stage-0 models. There are two stages all the while, First of all, we create meta-learning informational collections utilizing forecasts of the base model. Second, by utilizing a meta-learning set, and learn meta-models that can include expectations of the first model to the last conjecture.

1.3 Performance measurement parameters

These all the methods offer better performance outcomes for the measurement parameters such as Accuracy, Correlation, Root Mean Square Error, the coefficient of Determination are calculated as follows:

1.3.1 RMSE

Root mean square error is a mainstream equation to gauge the error rate of a regression show. Be that as it may, it must be contrasted with models whose errors are measured in a similar unit [41]. It is ascertained as takes after:

$$\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}(p_i - a_i)^2}}{n} \quad \text{.....} \quad \text{Eq(3)}$$

In Eq. (3) a is representing actual target, p is representing predicted target, along with total number of instances is n.

1.3.2 Correlation (r)

Correlation portrays the factual connections amongst genuine and anticipated values [41]. It is characterized as takes after:

$$correlation = \frac{\sum_{1=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{.....} \quad \text{Eq.(4)}$$

Where x representing to the actual value, y representing to the predicted value, $\bar{x}$ is representing to the mean of the every actual value, $\bar{y}$is the mean of the every one of predicted values, and n representing the quantity of instances. Correlation lies in [0, 1] and thought to be great if its value inclines toward 1.

1.3.3 Coefficient of determination (R^2)

The coefficient of determination (R^2) abridges the illustrative energy of the Regression display. R^2 depicts the extent of change of the reliant variable clarified by the regression show. In the event that the regression model is impeccable then R^2 is 1 and on the off chance that the regression model is an aggregate disappointment then R^2 is zero i.e. no change is clarified by Regression. The Coefficient of Determination is figured by taking the square of r (i.e. correlation) [41]. It is characterized as takes after:

$$R^2 = r * r. \quad \text{.....} \quad \text{Eq.(5)}$$

1.3.4 Accuracy

The accuracy is figured as rate deviation of predicted target value by means of actual target value with worthy error [41].

$$Acc = \frac{100}{n}\sum_{i=1}^{n} q_i \quad \text{.....} \qquad \text{Eq.(6)}$$

$$q_{i=}\begin{cases}1 & if\ abs(p_i - a_i) \leq err \\ 0 & Otherwise\end{cases}$$

Where a representing the actual target, p representing predicted target, err is the acceptable error, acc represents the accuracy and n is the total number of instances.

1.4 Classification Algorithms

There is much class of classification algorithms. Few of them are explained below.

1.4.1 Naive Baye's

It is the statistical classification algorithms which predict the probability of given instances. It follows Bayes's rule and assumes that variables are not dependent on each other in a given class called as conditional independence [29]. It has exhibited high performance for large databases.

1.4.2 Random Forest

It is a popular ensemble based classification algorithm which builds a randomized decision tree in bagging algorithm and produces excellent predictors [29].

1.4.3 AdaBoost

Adaptive boosting used to enhance the accuracy of classification algorithms. In adaptive boosting, weights are given to each training instances and after that classification algorithm is applied [29]. It is fast and can be accelerated by weight pruning.

1.4.4 SVM

Support vector machine uses the principle of risk minimization principle. This principle divides data or information into classes with maximum margin among classes [29, 42]. It is capable of learning in sparse, high dimensionality spaces with training samples to minimize error rate and complexity of classification algorithm.

1.4.5 Logistic Regression:

To estimate the discrete value i.e. binary number/results 0 or 1, yes or no, true or false, depends upon given sets of independent input variables we use logistic regression. In other words, we can say logistic regression predicts the possibilities of accidents of an event or incidents by fitting to logit function, so we also called logit regression, since it predicts the probability and output results lie between o and 1 as expected.

Chapter 2: Literature Survey

In the following chapter, the analysis is performed on Machine Learning, ensembling, and various diversity ensemble methods and techniques. Below are abstracts of those analyses performed by various research fellows.

2 Machine Learning

Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics [20]. Machine Learning is a branch of science that arrangements with programming the frameworks such that they naturally learn and enhance with understanding. Here, learning implies perceiving and understanding the info information and settling on insightful choices in light of the provided information [16, 18].

It is exceptionally hard to cover every one of the choices in view of every single conceivable information. To handle this issue, algorithms are created. These algorithms fabricate information from particular information and past involvement with the standards of likelihood hypothesis, rationale, measurements, fortification learning, look, combinatorial optimization, and control hypothesis [16, 18].

There are a few approaches to execute Machine Learning procedures, be that as it may, the most ordinarily utilized ones are regulated and unsupervised learning.

2.1.1 Supervised Learning

Supervised learning manages to learn a capacity from accessible preparing information. A supervised learning algorithm breaks down the preparation information and produces a construed work, which can be utilized for mapping new illustrations [16, 18]. Supervised learning as relapse (for ceaseless yields) and classification (for distinct yields) is an essential element of insights and Machine Learning, also for examination of informational indexes or as a sub goal of a more difficult issue [19].

Here some examples are shown below

i) Classifying e-mails as spam

ii) Labeling web pages based on their content

iii) Voice recognition.

There are many supervised learning algorithms, for example, neural networks (NN), Support Vector Machines (SVMs), as well as Naive Bayes classifiers.

2.1.2 Unsupervised Learning

Unsupervised learning understands unlabeled information without having any predefined dataset for its preparation. Unsupervised learning is a to a great degree effective instrument for breaking down accessible information and search for examples and patterns. It is most usually utilized for bunching comparative contribution to consistent gatherings. Basic ways to deal with unsupervised learning incorporate [16].

i) k-means
ii) self-organizing maps
iii) hierarchical clustering

2.1.3 Semi-supervised learning

To defeat the drawback of supervised learning algorithms with the intention, they can't make utilization of unlabeled information, semi-supervised learning (SSL) has been anticipated to use both marked and unlabeled information [17]. Common approaches to semi-supervised learning include-

i) Generative Models
ii) S3VMs
iii) Graph-Based Algorithms
iv) Multi-view Algorithms
v) Self Training

2.2 Application of Machine Learning

In the era of science, the Machine Learning works in various domain like Research, Biology, Statistics, Big data Analytics, Data Mining, Medical Science, Robotics, etc. Some of them are shown in Table 2.1.

Table 2.1 Application of ML

S.no	Paper Title	Description	References no
1	Machine Learning for the Detection of Oil Spills in Satellite Radar Images	They described Machine Learning application to major environmental problem discovery of oil spills from radar images on the top of sea's surface. They also covered the application cycle from the problem formulation stages to the delivery of a system for field testing.	[21.
2	An Introduction to MCMC for Machine Learning	The purpose of this paper are. It introduced the Monte Carlo technique among emphasis on probabilistic Machine Learning. It reviews the core building blocks of modern Markov chain.Monte Carlo simulation and Lastly, it discusses new interesting research horizons.	[22]
3	Gaussian Processes in Machine Learning	They give a basic introduction to the Gaussian process regression. Understanding and understanding the responsibility of the stochastic process and it is used to define distribution on the tasks. They presented simple equations for incorporate training data and learning the hypersaparatometer using marginal probabilities. They explained the practical compensation of Gaussian Process and a look at the present trends in GP work.	[23]
4	The Discipline of Machine Learning	Discussed Machine Learning is its significant real-world applications, listed below Speech recognition Computer visualization Bio-surveillance automaton control Accelerating experimental sciences.	[24]
5	Deep learning applications and challenges in big data analytics	This paper focuses on ML techniques in the perspective of big data and modern computing environments. Specifically, we aim to investigate opportunities and challenges of ML on big data. Big data presents new opportunities for ML. For instance, big data enables pattern learning at multi-granularity and diversity, from multiple views in an inherently parallel fashion. In addition, big data provides opportunities to make causality inference based on chains of the sequence. Nevertheless, big data also introduces major challenges to ML such as high data dimensionality, model scalability, distributed computing, adaptability, and usability.	[20]

Below is a list of some more applications of Machine Learning [27].

2.2.1. Automating Employee Access Control

Amazon, one of the guide of machine-learning based proposal recommendation engines and value segregation calculations, propelled a Machine Learning challenge on Kaggle to decide if it was conceivable to computerize worker get to giving and disavowal. Amazon has an impressive dataset of worker parts and representative gets to levels. They're attempting to build up a PC calculation that will anticipate which representatives ought to be conceded access to what assets. As indicated by Amazon, "This auto-get to models look to limit the human inclusion required to concede or renounce representative get to."

2.2.2. Protecting Animals

Cornell University is chipping away at an algorithm to distinguish whales in the sea in light of sound recordings with the goal that boats can abstain from hitting them. Additionally, Oregon State University is taking a shot at programming that will figure out which winged creature species are on a given sound recording gathered in field conditions.

2.2.3. Predicting Emergency Room Wait Times

Healthtech organizations and social insurance associations are utilizing a procedure called Discrete Event Simulation to foresee sit tight circumstances for patients in the crisis office holding up rooms. The models utilize variables, for example, staffing levels, tolerant information, crisis office outlines, and even the design of the crisis room itself to anticipate hold up times.

2.2.4. Identifying Heart Failure

IBM specialists have figured out how to concentrate heart disappointment analysis criteria from free-content doctor notes. They built up a Machine Learning algorithm that sifts through doctors freestyle content notes (in the electronic wellbeing records) and integrates the content utilizing a method called "Natural Language Processing" (NLP).

Like the way a cardiologist can read through another doctor's notes and make sense of whether a patient has heart disappointment, PCs can now do likewise.

2.2.5. Predicting Strokes and Seizures

Singapore-based startup Healing propelled an application called JustShakeIt that empowers a client to send an urgent caution to urgent links and/or parental figures essentially by trembling the telephone with a single hand. The program utilizes a Machine Learning algorithm to recognize actual urgent situation shakes and daily jostling. Notwithstanding the JustShakeIt application, Healing is taking a shot at a model that breaks down patients' PDA accelerometer information to help distinguish cautioning signs for ceaseless neurological conditions.

2.2.6. Predicting Hospital Readmissions

My own particular startup, Additive Analytics, is taking a shot at a Machine Learning model that recognizes which patients are at high danger of readmission. Utilizing our exclusive prescient model, healing centers can foresee emergency room confirmations before they happen—enhancing care results and lessening costs.

2.3 Ensembling methods

2.3.1 Definition

Set of predictive model is created with the help ensemble method, which is method of learning and combine their output in the same forecast. The objective of combining multiple models together, to reach a better future performance, and in many cases, it is shown that ensembles can be more accurate than a model [28].

2.3.2 Ensemble Techniques

The use of various schemes for the construction and integration of base models, some techniques for ensembling are Voting, Bagging, Boosting, Random Forest, and Stacking.

2.3.2.1 Voting

Entirely, Voting isn’t considered as method of ensemble, but there is a way to get involved in the base model, it’s not worried about the era of the hinge models. Still, we incorporate it in this determination of ensemble models since it can be utilized for joining

models paying little respect to how these models have been built. As said sometime recently, voting joins the expectations of support models as indicated by a fixed voting plan, which doesn't rely on upon the learning information. It thinks about to taking the straight blend of the models. The least complex sort of voting is majority vote (likewise called lion's share vote), where every base model makes a choice for its forecast. The expectation that gathers the majority votes is the last forecast of the ensemble. On the off chance that we are foreseeing a numeric esteem, the ensemble forecast is the normal of the expectations of the base models [28]. A fascinating part of voting is that, as a result of its straightforwardness, it takes into consideration some hypothetical investigation of its effectiveness. For instance, after modeling binary issue (an issue with both conceivable esteems, e. g., +ive and -ive) it has been3628shown that, if we have a gathering of independent base models than each with achievement likelihood (exactness) more prominent than 1/2, i. e., superior to anything arbitrary speculating, the exactness of the ensemble increments as the quantity of models increments [28, 29, 30, 31].

2.3.2.2 Bagging

Bagging (short for bootstrap accumulation) [32] is a voting method where construct models are found out in light of various variations of the learning informational index which are created through bootstrapping (kind of bootstrap testing) [33]. Bagging must be utilized simultaneously with an unsteady learning algorithm (e. g., choice trees or neural systems), where little changes during the knowledge set to bring about to a great extent extraordinary classifiers [28]. A novel intrusion area framework in perspective of outfit technique for Machine Learning is proposed. The Bagging technique for the outfit with REPTree as the base class is used to realize intrusion disclosure structure [38].

2.3.2.4 Boosting

Boosting [34] 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 methodologies is that during bagging the integrally of the built base models is absent to risk, while inside boosting we attempt to create reciprocal base models by learning resulting models, considering the missteps of past models. The methodology begins by learning the respectable starting

point model lying on the whole learning set with similarly weighted cases. For the following base models, we need them to effectively anticipate the cases that have not been accurately anticipated by past base models. Subsequently, we increment the weights of these illustrations (or reduction the weights of effectively anticipated cases) and take in another base model. We quit adapting novel base models while some stopping standard is fulfilled.

2.3.2.5 Stacking

Stacking [37] 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 known as meta (or stage 1) model is found out and utilized for consolidating base (or stage 0) models. The methodology has two stages. To start with, we make the meta-learning educational record via the desires of the base models. next, with the use of meta-learning set we take in the meta-display which can merge desires of base models hooked on the last estimate.

2.3 Diversity in Ensembles

Ensembles of classifiers have garnered great interest in recent years as it has been shown by several studies that, both theoretically and empirically, they can outperform single classifiers when the members of the ensemble are as accurate as possible and make few coincident errors. Since it is highly unlikely to train the perfect classifier that makes no errors, we need an ensemble of classifiers in which members make different errors to complement each other.

Robert. Schapire [1] focused on quality of powerless learnability of Machine Learning models with the utilization of learning algorithm in dispersion free Probably Approximately Correct (PAC) learning model. In this paper, they have proof based on filtering of distribution the weak learning algorithms to eventually learn nearly to the entire distribution.

Ludmila I. Kunchena [2] discussed about the procedures of contrasts in classifier outfits and their affiliation with troupe accuracy. They perform four pairwise measures and six

non-pairwise measures, in four experiments they show designed to determine the relationship between accuracy of the team and measures of the diversity (Some measures are taken directly and some from literature).

Eric Bauer and Ron Kohavi [4] introduced the method for voting classification algorithm like Boosting, Bagging and AdaBoost by using the following method they are successful to improvement in accuracy of some classifier for real-world datasets. They show boosting algorithm were generally better than bagging in some datasets used by them. The positive correlation between increases in the average tree size in Adaboost & its success in reducing in error rate.

Louisa Lan [3] discussed implementation and theoretical issues while performing ensembling with different classifiers. They used different topologies (conditional, hierarchical, and hybrid etc) to perform categorization of ensemble methods.

Padraig and John Carney [5] conclude that feature subset selection is useful techniques to create diversity in ensembles. They disagree that the diversity needs near examine in the formation of the combinators or ensembles. In this paper, they were present an technique to ensemble making and suggest entropy along with cross-entropy as dealings of diversity.

Guerra-Salcedo and Darrel Whitley [6] worked on feature subset selection id used as a system for introducing variety in ensembles of Machine Learning models. They compare Boosting as well as Bagging techniques using three diverse approaches for feature selection and ensemble construction. An another important contribution of the paper is the modified boosting scheme (results labeled in boost), which has empirically shown to be more effective than traditional Boosting.

Dariusz Brzezinski and Jerzy Stefanowski [7] discussed in the perspective of fix data, aims of this dissertation is summerize the impact of concepts flow on diversity measures intended for streaming ensembles. They used six main diversity measures be-Disagreement (D), Kohavi Wolpert (KW), Double fault (DF), Interrater agreement (k), Yule's Q statistic (Q), and Coincident failure diversity (CDF) to modify their calculation to data stream requirements.

Shuo Wang *et al.* [8] this paper aims to get a deeper perceptive if ensemble diversity has a positive effect on the taxonomy of imbalanced data sets. They explained when and why diversity measured through the Q-statistics, they bring two classification pattern based on the accuracy anticipated by Kuncheva et al. [9].They have also used six measures of diversity[7].They find the strong correlation between diversity and performance measures. Some results illustrate the positive impact of diversity in marginal class.

When they have studied for class imbalance learning got a relationship between ensemble diversity and performance measures. To complete understanding of G-mean & area under the curve (AUC) they used Q-statistics as the diversity measures including three functions- recall, F-measures, and precession.Some mathematical relation between Q-statistics and the single class measures are discussed by Kuncheva[9]

The combination or union of the results of some classifiers is only skilled if combinations are disagreeing on some internal inputs [10]. They mention the assess of deviation as to the diversity / doubt of the ensemble. on behalf of regression evils, this means that the squared error is usually used to compute accuracy, and variation is used in the direction of variation measure. In this setting [12], demonstrates that the speculation blunder E, of the group, can be communicated as E = $\overline{E} - \overline{D}$; where $\overline{E}$ is mean error and $\overline{D}$ is the diversity of the ensemble. This outcome infers that expanding ensemble diversity even as keeping up the normal blunder of ensemble individuals ought to prompt a reduction in ensemble errors. Dissimilar to relapse, for the grouping undertaking the, above basic straight affiliation doesn't hold among E, $\overline{E}$ and $\overline{D}$. Be that as it may, there is as yet solid motivation to trust that expanding diversity ought to abatement ensemble errors [11].

Present have been a couple measures of diversity meant for classifier outfits anticipated in the composition. During a current review [2] thought about ten unique diversity measures. They found that a large portion of these measures are very connected. Be that as it may, to the best of our insight, there has not been an indisputable review indicating which compute of diversity is the most excellent to use in favor of developing and assessing ensembles.

Prem Melville and Raymond J. Mooney [13] represent a new technique for generate ensembles called DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples). When we use the DECORATE method for small datasets it obtains higher accuracy as compared to Boosting.

Robert E. Banfield, *et al.* [14] have introduced a new technique called 'Thinning'.It is used as diversity metrics and means for modifying the diversity of ensembles.The thinning also show the greater correlation in diversity value and increase in ensemble accuracy.They have also created a different thinning technique based on exactness of ensemble and classified with regards to the training datasets.

Gabriele Zenobi, Padraig Cunningham [15] On the basis of various feature subsets, focused on the pieces of classifier and describes an algorithm which select special feature sets (and thus members of the ensemble cast) rather than reducing individual error, rather To maximize ambiguity. They present a study on pieces of K-nearest neighbor (KNN) classifier, who are skilled on three diverse datasets among two mountain climbing approach. The outcome show that the method has overcome the tactics based on error just on the basis of error.

Zhang et al. [48] demonstrated that an ensemble of three identical classifiers with 95% accuracy is worse than an ensemble of three classifiers with 67% accuracy and least pairwise correlated error. Due to this reason, diversity among the ensemble members is crucial for having fewer prediction errors.

There are several methods proposed in the literature to generate a diverse set of classifiers. These studies can be divided into two categories based on how they create diversity in an ensemble: implicitly and explicitly. Methods such as Bagging [42], Boosting [51], and Random Forest [44] introduce diversity by sub-sampling or re-weighting the training instances for the base classifiers. Two problems with these methods are that it is hard to decide how diverse the classifiers are and what the ensemble size should be. Due to this reason, measuring diversity and controlling it to increase the performance of an ensemble have become important. To answer the first question, several diversity measures have been proposed, some of which are borrowed from statistics,

including disagreement measure, Q-statistics, and double fault. Some of the diversity measures proposed in the literature are Kohavi-Wolpert [49], generalized diversity [47], interrater agreement [43], measure of difficulty [52], conditional double fault [48], etc.

Margineantu and Dietterich [50] suggested a Kappa-error plot for which diversity of every pair of classifiers in an ensemble is plotted against average of the individual errors of the two classifiers. This plot showed that best pairs are the ones with low errors and high diversity. However, the shape of the curve also revealed that there is a trade-off between the accuracy of the pair and its diversity. So, the success of ensemble pruning methods lies in balanced accuracy/diversity trade-off, where choosing only the most diverse classifiers or the most accurate individual classifiers to form the sub-ensemble decreases the generalization capability.

Melville and Mooney [53] proposed a method called DECORATE which increases diversity of an ensemble by making use of artificial data to train the base classifiers. At each iteration, DECORATE generates artificial data based on the performance of the current ensemble's response. More specifically, the labels for these artificially generated training instances are chosen so as to differ maximally from the current ensemble's predictions. Then artificial data is added to the original training data to train a new classifier. They compared the diversity with the ensemble error reduction, i.e., the difference between the average error of the ensemble members and the error of the entire ensemble by computing Spearman's rank correlation between the two. It was claimed that the fairly strong correlation between the two is another indication for increasing ensemble diversity to reduce generalization error.

Later, Tang et al. [45] presented an in-depth analysis of six of the diversity measures discussed in [46] and showed their relationship with the minimum margin of an ensemble. They analyzed how generalization of an ensemble changed with respect to diversity when the average accuracy of the base classifiers is fixed, and how diversity and the average base classifier accuracy interact with each other. Based on the experiments, the claim was that exploiting diversity measures to seek diversity explicitly is ineffective. First, the change of measured diversity cannot provide consistent guidance on whether a set of base classifiers has good generalization performance. Second, the diversity

measures are negatively correlated to the average accuracy of the base classifiers. Finally, it was shown that if the average accuracy of the base classifiers is regarded as a constant and the maximum diversity is achievable, maximizing the diversity among the base classifiers is equivalent to maximizing the minimum margin of the ensemble on the training samples. Since the true distribution of the data is unknown, it was suggested that like SVM a diversity measure should contain a regularization term and all of the discussed diversity measures contain no regularization term. Therefore, even if these existing diversity measures can be maximized, the achieved ensemble may overfit.

Kapp et al. [54] also investigated the relationship between diversity and margin theory 21 from the perspective of which margin definition is used. There are three main measures related to margin theory: minimum margin, cumulative margin distributions, and average margin. They show that individual performances of the base classifiers is one factor that contributes to the overall ensemble performance, but it is not sufficient. Thus, some diversity is needed to get the highest majority vote performance.

Experimental results showed that the average margin is stable and minimum margin is unstable. But, maximizing average margin chooses the ensembles with the strongest individual members in a given pool. This leads to low diversity as the base classifiers chosen in this manner will be very similar. Rather than using average margin, it is recommended using Chebyshev's inequality (CI) which states that probability of error should be less than the variance of margins divided by average margin squared. This implies that the ensembles must be sufficiently confident on their decisions with a certain majority vote, at least 50% in average.

De Olivieria et al. [58] addressed the accuracy/diversity dilemma for heterogenous ensembles using single and multi-objective GA approaches for analyzing accuracy and diversity separately, and jointly. The Q-statistic is used to measure diversity of the ensemble. It was concluded that similar to homogeneous systems the combination of diversity and accuracy can lead to more accurate ensemble systems, when compared with these two parameters individually. Even though these studies attempted to show the relationship between accuracy and diversity and to answer how much diversity is enough, these questions are not completely answered.

Hsu and Srivasta [55] changed the direction of the focus and analyzed the relationship between diversity and correlation of the classifiers instead of ensemble accuracy. The 22 relationship between diversity and correlation of the classifiers in ensembles was analyzed theoretically and formulated in a nonlinear function. They were able to derive a critical value for disagreement measure. It was shown that before the critical value, higher diversity reduces correlation which is usually associated with a better ensemble. When diversity crosses the critical point, increasing diversity increases the correlation while highly correlated classifiers usually correspond to an inferior ensemble.

Brown [57] investigated the relationship between accuracy and diversity by decomposing the ensemble's mutual information into accuracy and diversity terms and showed that the diversity of an ensemble exists at multiple orders of correlation. However, estimating the interaction of these multiple correlations is complicated and there is no proposal for applying this in practice. Later good and bad diversity concepts were introduced, and their relationship with the upper/lower limits were defined on majority voting error by Brown and Kuncheva.

Brown and Kuncheva [56] Here, the majority vote error is divided into three components: average individual accuracy, "good" diversity and "bad" diversity. The two diversity terms are related to the majority vote limits. Good diversity is derived to be the number of incorrect votes when the ensemble is correct. On the other hand, bad diversity is the number of correct votes when the ensemble is incorrect. They argued that a diversity measure should be naturally derived as a direct consequence of two factors: the loss function and the combiner function. It is recommended to construct algorithms like DECORATE as using artificially constructed data examples produces diversity in majority voting by making the individuals disagree wherever possible with the ensemble.

2.5 Research Gaps:

This section tells about the gaps encountered during the research by reviewing the already existing literature in the area of Ensembling problem in machine learning and odor prediction of the chemical.

i) forecasting the odor of a molecule is still very demanding. Fragrance chemists integrate many molecules to acquire a new constituent, but among the newly acquired constituent, most of them will not have the qualities chemists are craving for and are required to be experienced by subjects. This is also called the stimulus-percept problem [35].

ii) The variety of ensembles is a pivotal problem in machine learning, which has proved to be an unfolding research area in past few years in the field of machine learning. Classification algorithms are troubled by the variety and variability of problem for a particular dataset that is established on regression [28]

iii) Security and privacy concerns being the major problems in huge amounts of data. Here data will be first examined and then excavate for particular order or arrangement called pattern [17, 49].

iv) Scalability and complexity are major demanding complications in the domain of machine learning. Traditional tools are not able to control and manage large datasets [12, 49].

v) Timeliness is an additional concern for large datasets in machine learning. The size of dataset increases with the increase in time taken during analysis [12, 49].

2.6 Problem Formulation

The diversity in ensembles is a crucial problem in machine learning, which has become an emerging research region in modern time. Classification algorithms afflicted by the diversity problem for a certain datasets taken in this construction. There exists problem where the data sets are so large and very small that it isn't possible to become skilled at a model on the whole data set. It is sometimes more efficient approach is to separation the data into smaller parts, be trained one model for all part, and merge the outputs of these models hooked on a single prediction. when modeling a binary problem (a problem with two probable values, e. g., optimistic and pessimistic) it has illustrate, if we have an

ensemble piece of equipment with independent base support models each with success result probability (accuracy) greater than 1/2, i. e., better than random guessing, the accuracy of the ensemble increases as the number of base models increases [28].

Olfaction (the sense of smell) is the slightest implicit of the five senses. We utilize it frequently in our daily lives choosing foodstuff that is not spoiled, as an early-warning indication of a gas reveal or a fire, and in the enjoyment of scent and wine. Although many centuries of thought concerning the fundamental mechanisms of smell, predicting the smell of a molecule is still difficult. Fragrance chemists produce many molecules to acquire a fresh ingredient, but the majority of these will not have the preferred qualities and need to be experienced by subjects. This is known as the stimulus-percept problem [35], In order to tackle the olfaction stimulus-percept problem, we will take improvement of the largest existing smell-testing study to date, to evaluate algorithmic predictions of the smell that a molecule produces from its physical and chemical features. In this construction, we try to explore some machine learning methods(RF, NN, LM, Cubist, Enet, ICR, PCR, LR with SS) belongs to different-different categories(Tree, Linearity, Network etc) with physicochemical properties to predict the odor of a chemical compound.

2.7 Objectives

i) To learn and examine existing state of the art of machine learning algorithms, methods, techniques and models.

ii) To propose an ensemble machine learning framework to predict if certain chemical compounds smell will be liked by humans or not.

iii) To evaluate the performance of proposed framework using a various parameter like accuracy.

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

Chapter 3: Proposed Diversity Ensemble Framework (DEF)

3.1 Workflow of Diversity Ensemble

To make experiments more interesting the first part is to select new imbalanced data sets that have different attributes, different instances, and different parameters. The second part has preprocessed the data which include data format adaptation and data sampling. In data format adaptation first formats of datasets must be converted into.CSV files which are required by R interface. Data sampling depends on two parameters - percentage and bias. Below Figure 3.1 shows the workflow of Diversity ensemble framework.

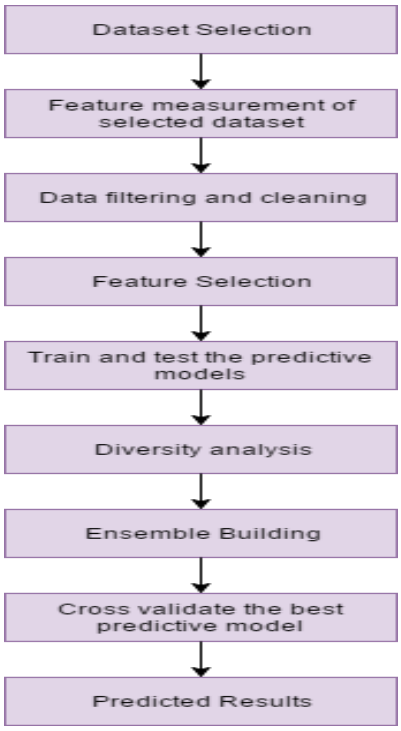

Figure 3.1 Work flow of DEF

We perform ensembling on various Machine Learning methods such as Random Forest, Neural Network, Linear Model, Cubist, Elastic net, Independent Component Regression, Principal Component Regression, Linear Regression with Stepwise Selection, on each data sets under following measures: Accuracy, Root mean square error, Coefficient of

Determination, and Coefficient of Correlation and we compare obtained results to determine best Machine Learning algorithm for each dataset. The research begins by selection of the imbalanced datasets. An imbalanced dataset is in which classes are not represented uniformly. These kinds of datasets are composed of typically two categories: Majority class and Minority class.

3.2 Parallel ensemble architecture

If base classifiers were all identical, there would be no need to build an ensemble out of them, as the ensemble decision would be exactly the same as each single base classifier. There are two main ways of creating different classifiers: we can either train the same classifier algorithm on different training sets, or we can train different classifier algorithms on the same training set.

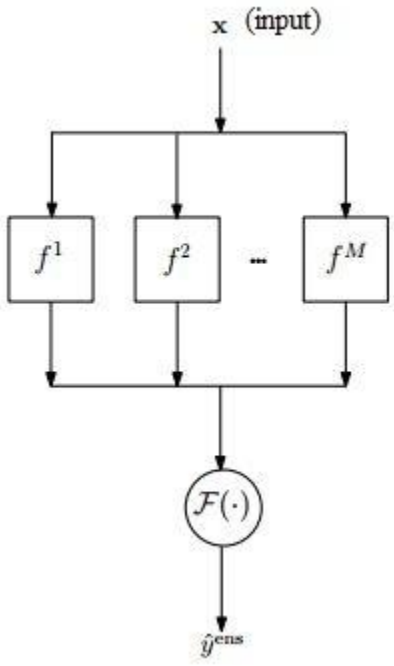

Figure 3.2: A parallel ensemble architecture

Here given an input dataset x to the base classifiers $f_1,...,f_M$, their outputs are combined according to some functional F(according to algorithm) to get the final ensemble decision $\hat{y}_{ens}$ as output.

3.3 Diversity Ensemble Framework (DEF)

Ensembles of classifiers have garnered great interest in recent years as it has been shown by several studies that, both theoretically and empirically, they can outperform single classifiers when the members of the ensemble are as accurate as possible and make few coincident errors. Since it is highly unlikely to train the perfect classifier that makes no errors, we need an ensemble of classifiers in which members make different errors to complement each other. The following element of DEF as shown in Figure (3.3) are discussed below-

3.3.1 Data Selection

Data selection is characterized as the way toward deciding the proper data sort and source, and in addition reasonable instruments to gather data. Data selection goes before the real routine with regards to data gathering. The essential goal of data selection is the assurance of fitting data sort, source, and instrument(s) that enable specialists to sufficiently answer look into inquiries. This assurance is regularly discipline-specific and is fundamentally determined by the way of the examination, existing writing, and availability to essential data sources. For this construction we taken the both dataset as input from the Dream Challenge and UCI Machine learning.

3.3.2 Data cleaning and filtering

A data set is a collection of data that describes attribute values (variables) of a number of real-world objects (units). With data that are technically correct, we understand a data set where each value. A data set is an accumulation of data that portrays characteristic esteems (factors) of various genuine items (units). With data that are in fact remedy, we comprehend a data set where each esteem. To begin with it can be specifically perceived as having a place with a specific variable, and second is put away in a data sort that speaks to the esteem space of this present reality variable. As such, for every unit, a content variable ought to be put away as content, a numeric variable as a number, et cetera, and this in a configuration that is predictable over the data set. Data filtering done in three steps Screening, Diagnosis and Editing as shown in Figure (3.3). Respectability issues can emerge when the choices to choose "fitting" data to gather are constructing

principally in light of cost and comfort contemplations as opposed to the capacity of data to enough answer examine questions.

3.3.3 Feature Selection

Feature selection is important step of removing immaterial and unnecessary features for the model construction [39]. Feature selection includes and excludes attribute present in the data without any revisions [43] [46]. Feature selection acts as a filter, as it mutes out those characteristics that aren't useful in addition to your existing features [45]. Feature selection enhances performance of classification algorithms. Feature selection methods are Filter method, Wrapper method, and embedded method [24] [41] [44]. These methods select those features which are best for performance of model. Filter method measure quality of important selected features, from machine learning algorithms, while wrapper methods needs application of classification algorithm to measure quality of selected features. For learning of optimal parameters, embedded methods perform feature selection [14] [31].

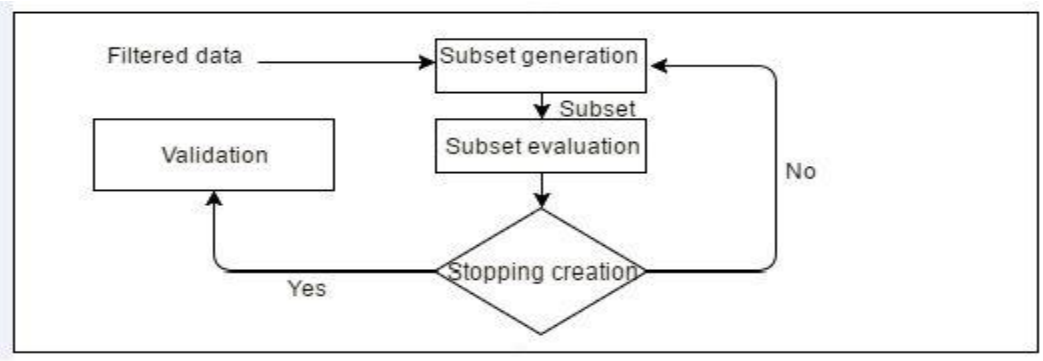

Fig.3.3 Feature selection

3.3.4 Training and testing of dataset

A preparation set is an arrangement of data used to find conceivably prescient connections. A test set is an arrangement of data used to survey the quality and utility of a prescient relationship. Test and preparing sets are utilized as a part of clever frameworks, machine learning, genetic programming, and statistics. During the implementation first step have to do preparation of the data then explore the data to use in predictive models

for final evaluation, and deliver the data for ensemble building. The following four steps are shown in Figure (3.4).

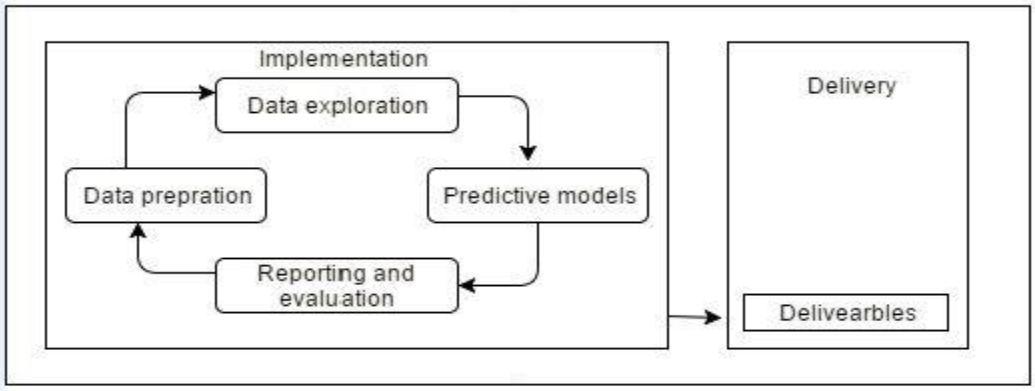

Figure 3.4 Training and testing

3.3.5 Ensemble building

Ensemble strategies are systems that make different models and then join them to deliver enhanced outcomes. An ensemble strategy as a rule delivers more precise arrangements than a solitary model would. This has been the situation in various machine learning rivalries, where the triumphant arrangements utilized ensemble techniques. Set of predictive model is created with the help ensemble method, which is method of learning and combine their output in the same forecast. The objective of combining multiple models together, to reach a better future performance, and in many cases, it is shown that ensembles can be more accurate than a model [28]. Ensembling define with the help of different machine learning algorithm as shown in Figure (3.5).

3.3.6 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.

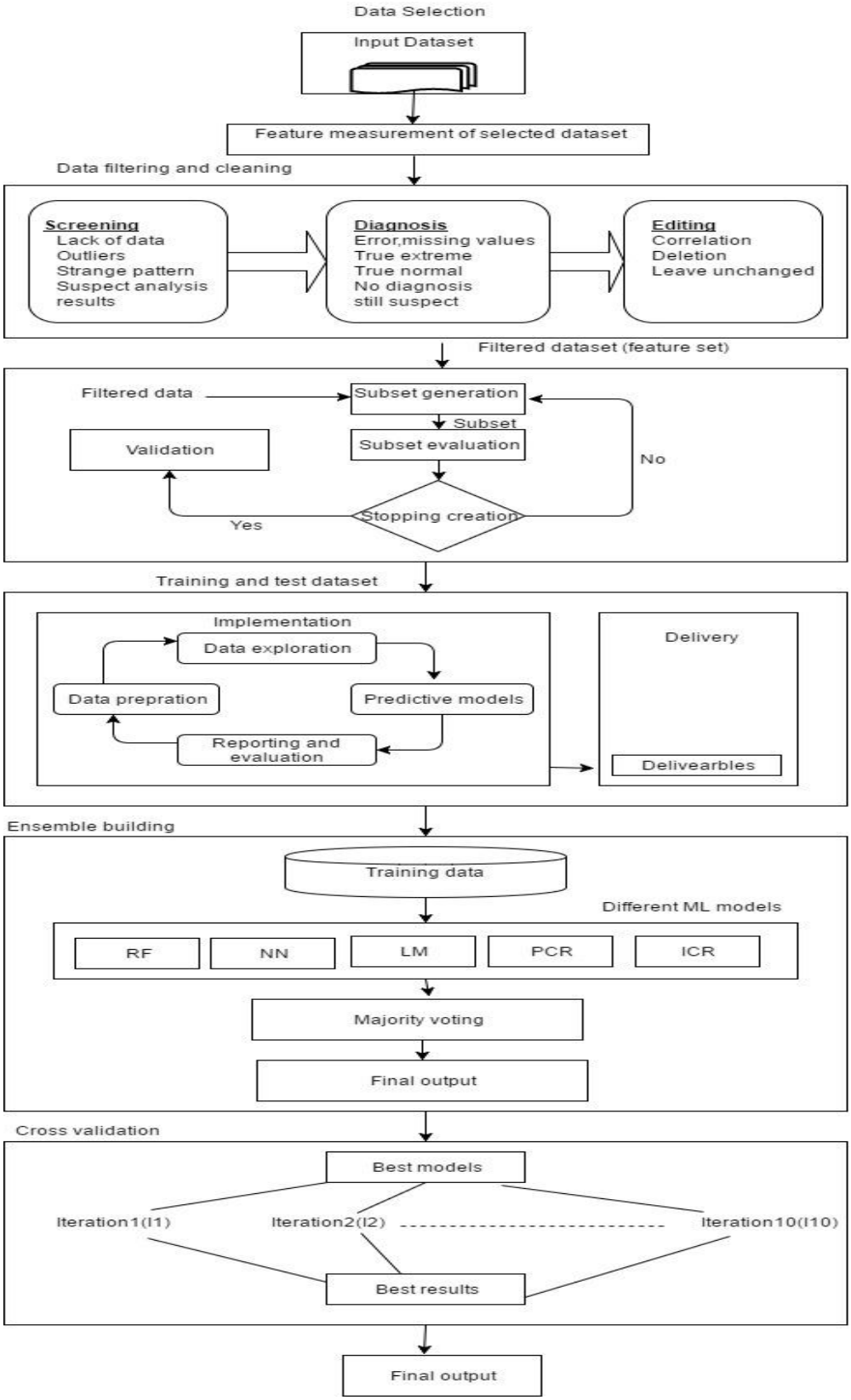

Fig. 3.4 Diversity ensemble framework

The concept of feature selection is the concept of selecting a subset of suitable features like predictors and variables. 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 eight Machine Learning methods for predicting, testing and training, namely:

i) Random Forest
ii) Cubist
iii) Neural Network
iv) Linear Model
v) Elastic net
vi) Principal component regression
vii) Linear regression with stepwise selection
viii) Independent component regression

A. Random Forest

Random Forest is one of the most popular machine learning algorithms which was created by Breiman [44]. 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 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 overfitting which is one of biggest problem in single decision tree. During training of model each

decision tree in model is trained on random subset of features of training data. In Bagging we select random sub samples of training data and train model on them but random forest is different then bagging as here we are not only choosing random sample of training data but we are choosing random sample of features as well. Random forest with help of multiple decision tree are much more generalized when compared to single decision tree as there are very less chance of overfitting. Random Forest can also be used to rank features. The idea of feature selection using random forest was given in the original paper of random forest itself.

B. Linear Model

Linear models portray a nonstop reaction variable as a component of at least one indicator factors. They can enable you to comprehend and anticipate the conduct of complex frameworks or examine test, money related, and organic information. Linear regression is a measurable strategy used to make a linear model.

C. Linear Regression with Stepwise selection

In statistics and measurements, stepwise regression is a technique for fitting regression models in which the decision of prescient factors is completed by a programmed system. In each progression, a variable is considered for expansion to or subtraction from the arrangement of illustrative factors in view of some pre-specified paradigm.

These all the methods offer better performance outcomes for the measurement parameters such as Accuracy, Correlation, Root Mean Square Error, the coefficient of Determination are calculated as follows:

i. RMSE

Root mean square error is a mainstream equation to gauge the error rate of a regression show. Be that as it may, it must be contrasted with models whose errors are computed in a similar unit [41]. It is ascertained as takes after:

$$\text{RMSE}= \frac{\sqrt{\sum_{i=1}^{n}(p_i-a_i)^2}}{n} \quad \text{.....Eq(7)}$$

In Eq. (7) is real target, p is predicted target, along with total number of instances is n.

ii. Correlation (r)

Correlation portrays the factual connections amongst genuine and anticipated values [41]. It is characterized as takes after:

$$correlation = \frac{\sum_{1=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{.....Eq.(8)}$$

Where x representing to the real value, y representing to the predicted value, $\bar{x}$is representing to the mean of the every actual value, $\bar{y}$ is representing the mean of the the entire predicted values, and n is representing the quantity of instance. Correlation deceit in [0, 1] and thought to be great if its value inclines in the direction of 1.

iii. Coefficient of determination (R^2)

The coefficient of determination (R^2) abridges the illustrative energy of the Regression display. R^2 Depicts the extent of change of the reliant variable clarified by the regression show. In the event that the regression model is impeccable then R^2 is 1 and on the off chance that the regression model is an aggregate disappointment then R^2 is zero i.e. no change is clarified by Regression. The Coefficient of Determination is figured by captivating the square the r [41]. It is characterized as takes after:

$$R^2 = r * r. \quad \text{.....Eq.(9)}$$

iv. Accuracy

The accuracy is figured as rate variation of predicted target value along with real target value with worthy error [41].

$$Acc = \frac{100}{n}\sum_{i=1}^{n} q_i \quad \text{..... Eq.(10)}$$

$$q_{i=}\begin{cases} 1 & if\ abs(p_i - a_i) \leq err \\ 0 & Otherwise \end{cases}$$

Where a representing the actual target, p representing predicted target, err is the adequate error, acc represents the accuracy and n is representing the total number of instances.

Once the dataset is balanced, important relevant features are selected from the largely balanced dataset. This allows better learning performance, better model interpretability,

and lower computational cost. Feature selection returns important features of the original ones, according to certain relevance evaluation criterion.

In our work the classifier's algorithms are given the different results in terms of accuracy, now we divided the all the classifiers into two part on the basis of their results(accuracy) one is the set of maximum accuracy models in strong bucket and second is minimum accuracy models in weak bucket as shown in Figure, when we used as single model for certain dataset then some methods gives the best results, while we Ensemble in group of two then see the two models they belong to weak bucket give the best result as compared to the strong bucket's ensembles. Here the diversity in ensemble comes into the picture. After taking one of the ensemble models with high accuracy from the combination of eight different models we applied K-cross validation on it to get better the performance of the proposed framework.

The basic technique is to train different classifiers like Random Forest, AdaBoost, and Linear Model on multiple subsets of the features and then ensemble all the model by the combination. After this, the resulted output is utilized for training and testing by different classification algorithms separately result are displayed in tables and graphs. The yield of all the eight algorithms is compared and analyzed and our result outperforms another state –of –the- art techniques.

Chapter 4: Design and Implementation of DEF

This chapter tells about the design and implementation performed during the research. Implementation details include installation of software, implementation of all the Machine Learning methods, algorithms and implementation of proposed framework named subset feature selection in Ensembling and also includes snapshots of entire implementation.

4.1 Design of DEF

Software design is a process to transform user requirements into some suitable form, which helps the programmer in software coding and implementation. Software design is the first step in SDLC (Software Design Life Cycle), which moves the concentration from problem domain to solution domain. It tries to specify how to fulfill the requirements mentioned in SRS.

4.1.1 Architectural Design - The architectural design is the highest abstract version of the system. It identifies the software as a system with many components interacting with each other. At this level, the designers get the idea of proposed solution domain. The architecture of the proposed framework has shown in figure (3.4) chapter 3.

4.2.1 Activity diagram

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.

The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc as shown in Figure (4.1). The overall details are discussed in previous chapter for each component of the activity diagram. The basic purpose of activity diagrams to captures the dynamic behavior of the system.

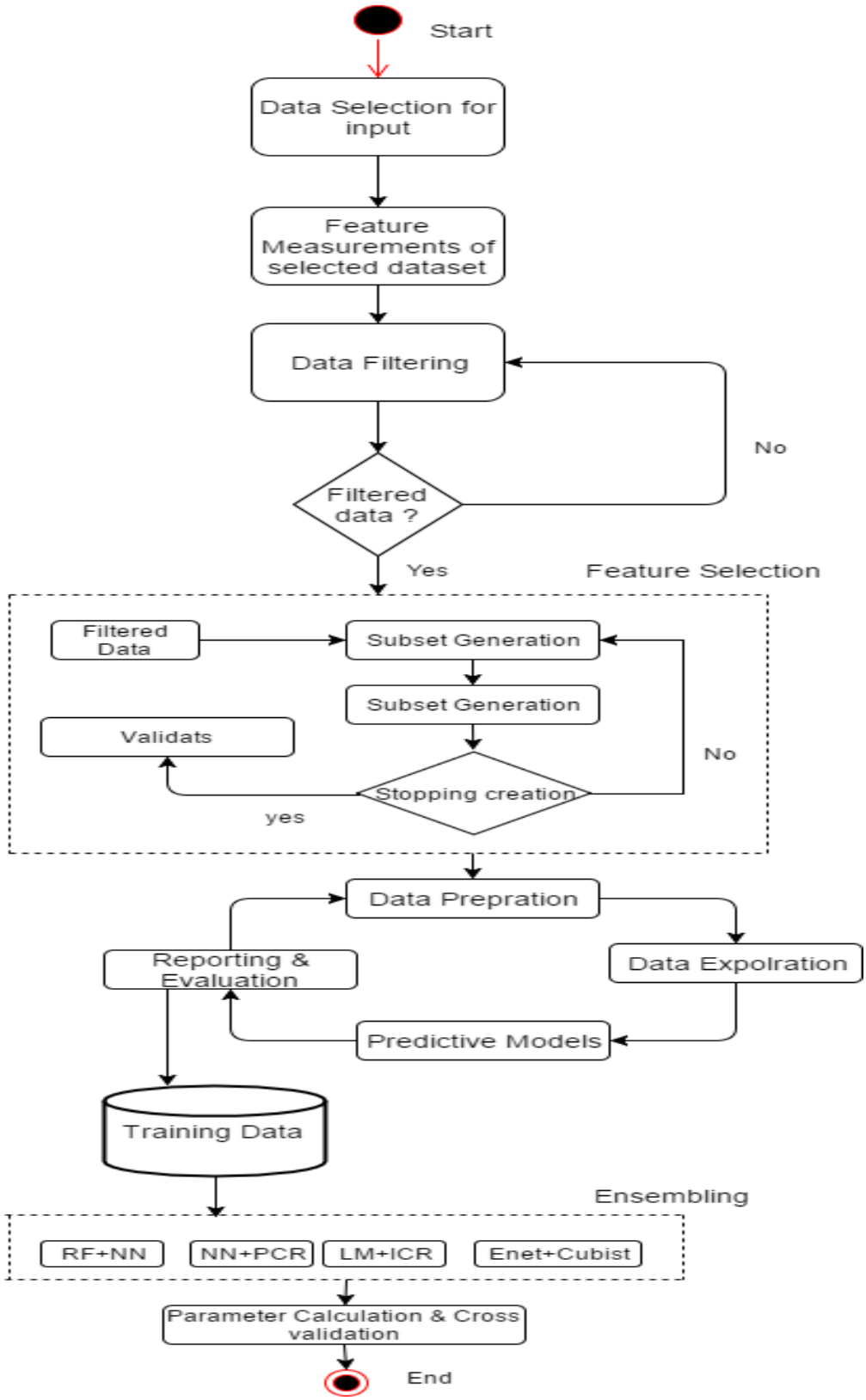

Figure 4.1 activity diagram for diversity ensemble

4.1.3 **Class Diagram**

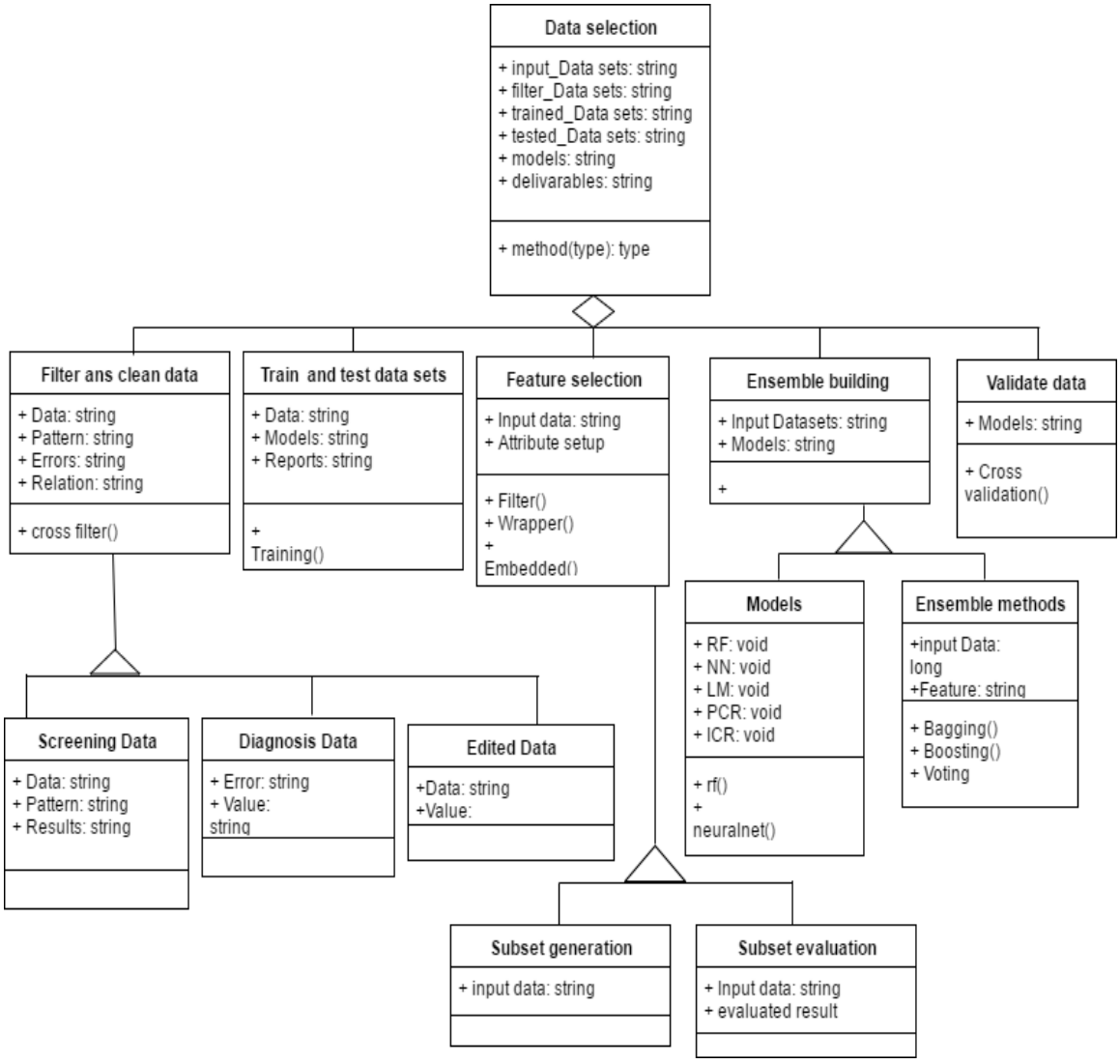

Fig.4.2 Class diagram for DEF

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. The discussion of all the classes of class diagram in previous chapter.

4.2 Experimental setup

4.2.1 Software and Hardware requirements (minimum)

Table 4.1 H/W and S/W Requirement (minimum)

1.	Processor	32 bit
2.	RAM	2 GB
3.	Hard Disk	80 GB
4.	Operating System	Windows 7
5.	Programming Language	R (Rattle)
6.	Platform	R Studio

4.2.2 Feature Selection Implementation

For a selection of features and subset of features from balanced datasets, we did all the implementation with the help of R language. R is installed in the system and Rattle is used to make a project where coding is done.

For a selection of feature and subset of feature selection, firstly we add the forest library. After adding library we include RF, Cubist, ICR, PCR, NN, LM, Enet, and leaps packages, after that, we performed read and write operation on balanced datasets. We have predicted the accuracy of each and every model of Machine Learning used in this construction.

4.2.3 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 libraries like caret, and boruta. After adding caret, and boruta, packages, we install random Forest package. After that, we performed read and write operation on the datasets and measure performance of the random forest in stipulations of accuracy. The random forest can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('randomForest')
library(randomForest)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="randomForest")
```

After creating a model we measure the performance of created model using in stipulations of accuracy and find the value of correlation, root mean square error, Rsqure are shown.

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

4.2.4 Neural Network Implementation

We did all the implementation with the help of R language. For Neural network implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install nnet package. After that, we performed read and write operation on the datasets and measure the performance of the Neural network in terms of accuracy. The neural network can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('neuralnet')
library(neuralnet)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="neuralnet")
```

After creating a model we measure the performance of created model using in terms of accuracy and find the value of correlation, root mean square error, Rsqure are shown

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

The same implementation has done for all remaining Machine Learning models used in this construction to calculate accuracy and find the diversity in ensembles by ensembling method, all the experimental results are shown in next chapter.

2.2.5 Linear Model Implementation

We did all the implementation with the help of R language. For Linear Model implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install lm package. After that, we performed read and write operation on the datasets and measure the performance of the Linear Model in terms of accuracy. The Linear Model can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('lm')
library(lm)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="lm")
```

After creating a model we measure the performance of created model using in the expressions of accuracy and find the value of correlation, root mean square error, Rsqure are shown.

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

4.2.6 Cubist

We did all the implementation with the help of R language. For Cubist Model implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install cubist package. After that, we performed read and write operation on the datasets and measure the performance of the Cubist Model in terms of accuracy. The Cubist Model can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('Cubist')
library(Cubist)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="cubist")
```

After creating a model we measure the performance of created model using in expressions of accuracy and find the value of correlation, root mean square error, Rsqure are shown.

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

4.2.7 Linear regression with stepwise selection

We did all the implementation with the help of R language. For Linear regression with stepwise selection Model implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install leaps package. After that, we performed read and write operation on the datasets and measure the performance of the linear regression with stepwise selection Model in terms of accuracy. The Linear regression with stepwise selection Model can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('leaps')
library(leaps)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="leapSeq")
```

After creating a model we measure the performance of created model using in expressions of accuracy and find the value of correlation, root mean square error, Rsqure are shown.

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

4.2.8 Principal component regression

We did all the implementation with the help of R language. For principal component regression model implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install pcr package. After that, we performed read and write operation on the datasets and measure the presentation of the principal component regression Model in provisos of accuracy. The PCR Model can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('pcr')
library(pls)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="pcr")
```

After creating a model we measure the performance of created model using in expressions of accuracy and find the value of correlation, root mean square error, Rsqure are shown

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

4.2.9 Independent component regression

We did all the implementation with the help of R language. For independent component regression model implementation Firstly we add various R packages and libraries like caret, and boruta. After adding caret, and boruta, packages, we install fastICA package. After that, we performed read and write operation on the datasets and measure the presentation of the independent component regression Model in provisos of accuracy. The ICR Model can be implemented using the following function which includes formula, trainDataset, and method.

```
install.packages('icr')
library(fastICA)
library(caret)
library(Boruta)
formula <- as.formula(paste(target, "~", paste(c(inputs), collapse = "+")))
model  <- train(formula, trainDataset,method="icr")
```

After creating a model we measure the performance of created model using in terms of accuracy and find the value of correlation, root mean square error, Rsqure are shown

```
#  Correlation
r <- cor(Actual,Predicted )
r <- round(r,2)
r

#  RSquare
R <- r * r
R <- round(R,2)
R

#  RMSE
rmse <- mean(abs(Actual-Predicted))
rmse <- round(rmse,2)
rmse

# Accuracy
accuracy <- mean(abs(Actual-Predicted) <=0.81)
accuracy <- round(accuracy,4) *100
accuracy
```

Chapter 5 Experimental Results And Analysis

To predict the intensity and pleasantness of the chemical in terms of accuracy, and also predict the diversity of the different Machine Learning models in model ensembling through accuracy. We choose a data set and apply 8 different Machine Learning model, the dataset belongs to regression data so we applied some Machine Learning model based on regression. The summary of datasets shown in Table (5.1)

Table 5.1 Dataset details

Dataset Name	**No. of Samples**	**No. of Features**
Dream Olfaction Prediction Challenge.	35345	24
Facebook Metrics Performance	501	19

5.1 Methodology

The methodology is divided into eight phases and each phase is described as:

Phase 1: In the first phase, huge unpublished data set based on wide-ranging smell-testing of 49 human subjects asked to snuffle 476 different smell chemicals will be taken.

Phase 2: The elimination of duplicates and absent value entries (data cleansing and filtering) from the dataset will be conceded out in the second phase.

Phase 3: Physicochemical parameters for each odor compound will be calculated in this phase.

Phase 4: In the fourth phase, a real-coded Self-adaptive Differential Evolution algorithm (SaDE) will be used to determine the significance of each attribute. Feature selection makes the forecast of model efficient, fast and truthful.

Phase 5: In the fifth phase, eight Machine Learning techniques: Random Forest, Neural Network, Linear Model, Cubist, Elastic net, Independent Component Regression,

Principal Component Regression, Linear Regression with Stepwise Selection would be used to train and test the dataset.

Phase 6: This is the evaluation phase where parameters such as Root Mean Square Error (RMSE), correlation, R2, and accuracy would be used to evaluate the efficiency of the method.

Phase 7: In phase seven, K-fold cross-validation would be used to compute the sturdiness of the best predictive techniques.

Phase 8: In last phase, we would get the results in terms of accuracy.

5.2 Results

An experimental result of proposed framework is presented below in Tables and Figures. The table contains outcome of 8 different Machine Learning models along with the value of Accuracy, Coefficient of Determination, Coefficient of Correlation**,** Root mean square error. The suggested framework is applied on Olfaction of chemical and to predict the Facebook metrics performance data sets As already explained that we have applied the regression algorithms on the dataset. These algorithms are as follows: Random Forest, Neural Network, Linear Model, Cubist, Elastic net, Independent Component Regression, Principal Component Regression, Linear Regression with Stepwise Selection. We have Compared the all these algorithms on the basis of their accuracy. The experimental results are shown below in the Table (5.2) and Figure (5.1).

Table 5.2 Single model's results for data set-1

Model no	**Model Name**	**R**	**R**	**RMSE**	**Accuracy**
1	RF	0.47	0.22	0.36	92.81
2	LM	0.03	0	0.97	74.65
3	E net	0.44	0.19	0.451	69.66
4	Cubist	0.45	0.2	0.447	69.63
5	LR with SS	0.38	0.14	0.463	64.47
6	ICR	0.24	0.06	0.487	63.27
7	PCR	0.17	003	0.493	55.69
8	NN	0.1	0.01	0.87	42.51

The results are showing that Random forest gave the best result for dataset with highest accuracy and Neural Network and Principal component regression given the worst results so we perform ensembling for getting batter performance.

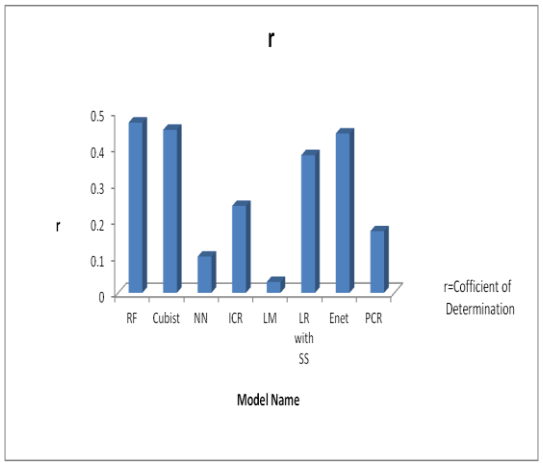

Fig.5.1 r plot

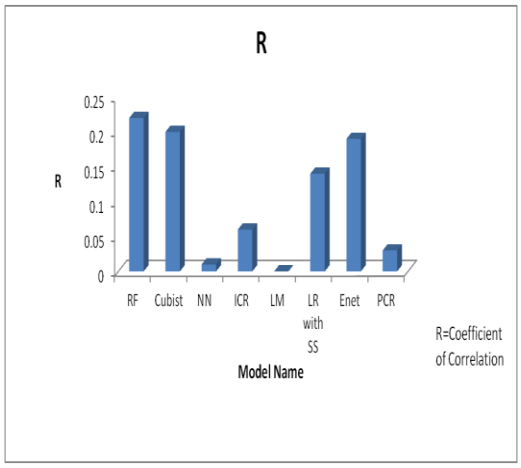

Fig.5.2 R plot

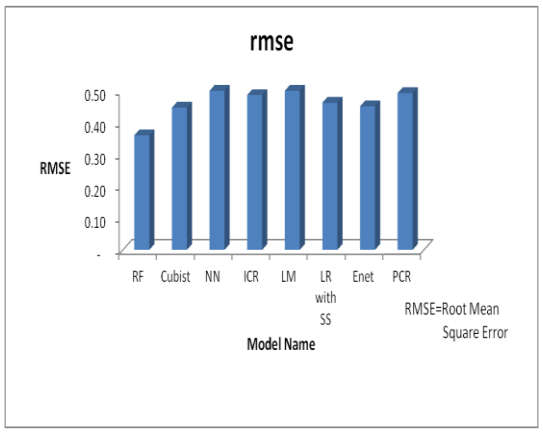

Fig.5.3 RMSE plot

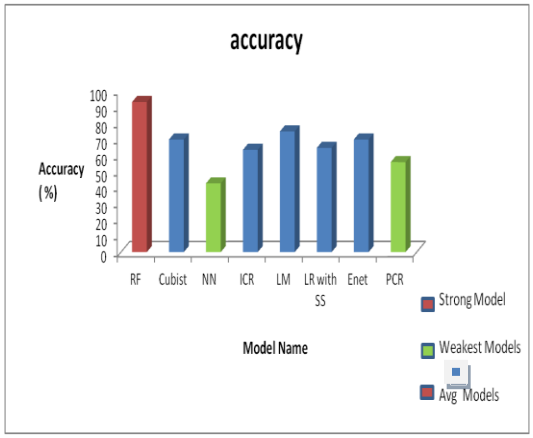

Fig.5.4 Accuracy plot

Here results of the all the performances measurement parameter like rmse, R square, R and accuracy graphs between different models are shown in Figure (5.1, 5.2, 5.3, 5.4) respectively.

Now to improve the accuracy and find diversity factor we have applied the ensembling algorithm, in which we combine two models and run for both the dataset. The improved experimental results are shown in Table (5.3).

Table 5.3 Ensembling Results for olfaction dataset

Model No	**Model Name**	R	R	RMSE	Accuracy
1	Enet+ ICR	0.37	0.14	0.47	89.62
2	PCR +Enet	0.37	0.14	0.47	90.62
3	PCR + ICR	0.25	0.06	0.48	94.81
4	PCR + ICR	0.25	0.06	0.48	94.81
5	PCR+LM	0.31	0.10	0.48	88.42
6	LR with SS +icr	0.37	0.14	0.47	84.83
7	PCR +LR with SS	0.70	0.49	0.38	90.02
8	Enet+LR_with_ss	0.41	0.17	0.46	80.64
9	LM+Enet	0.35	0.12	0.47	73.45
10	icr+lm	0.37	0.14	0.47	85.23
11	LR with SS +lm	0.42	0.18	0.45	79.64
12	NN+ LR with SS	0.34	0.12	0.47	97.21
13	NN+Enet	0.41	0.17	0.47	96.81
14	NN+LM	0.42	0.18	0.46	96.41
15	cubist +lm	0.63	0.40	0.39	86.23
16	ICR+RF	0.62	0.38	0.41	89.22
17	NN+PCR	0.15	0.02	0.50	97.80
18	PCR +Cubist	0.63	0.40	0.40	87.03
19	NN+Cubist	0.67	0.45	0.39	87.23
20	cubist +icr	0.65	0.42	0.38	88.02
21	cubist +LR with SS	0.65	0.42	0.38	88.02
22	RF+Enet	0.67	0.45	0.39	91.42
23	RF+Cubist	0.58	0.34	0.42	86.63
24	RF+PCR	0.58	0.34	0.42	86.63
25	RF+LM	0.59	0.35	0.41	84.03
26	RF+NN	0.63	0.35	0.41	84.00
27	ICR+NN	0.19	0.04	0.49	96.81
28	RF+LR_With_SS	0.63	0.40	0.40	88.82

After performing ensembling on the above models we got the highest accuracy of the ensemble no. 17 (NN+PCR) they are both gives worst performance when we use separately but after ensembling combination of two worst models gives the best results as display in Table 5.3.

Here the graphical representation of all the result got from the ensembling is shown:

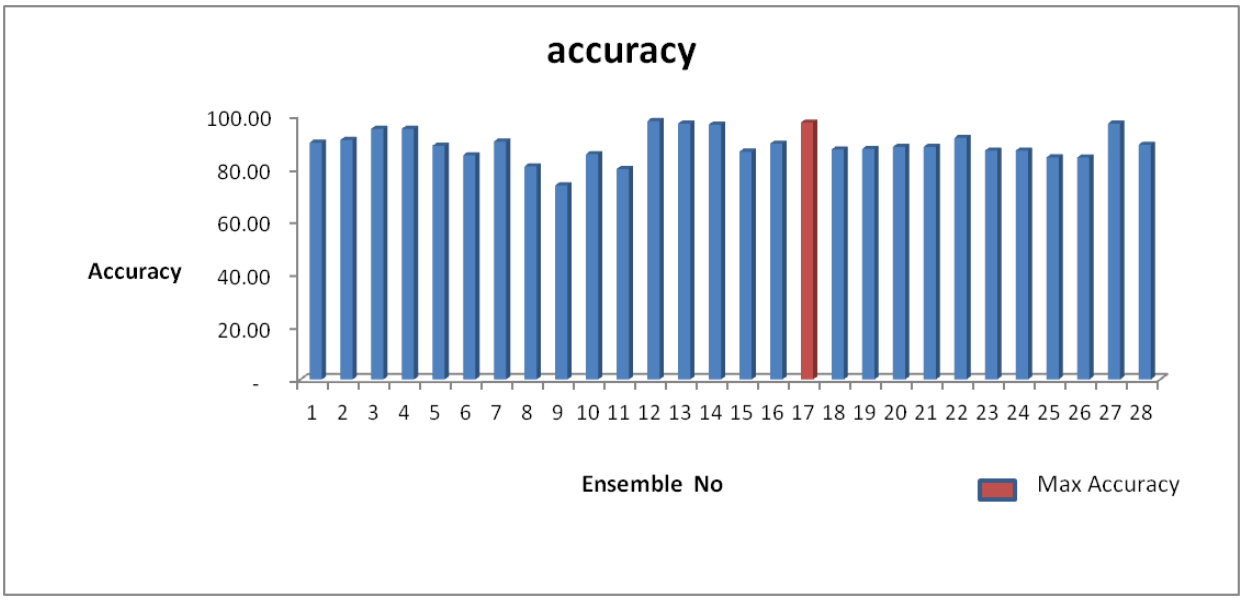

Fig.5.5 Accuracy plot with ensembling

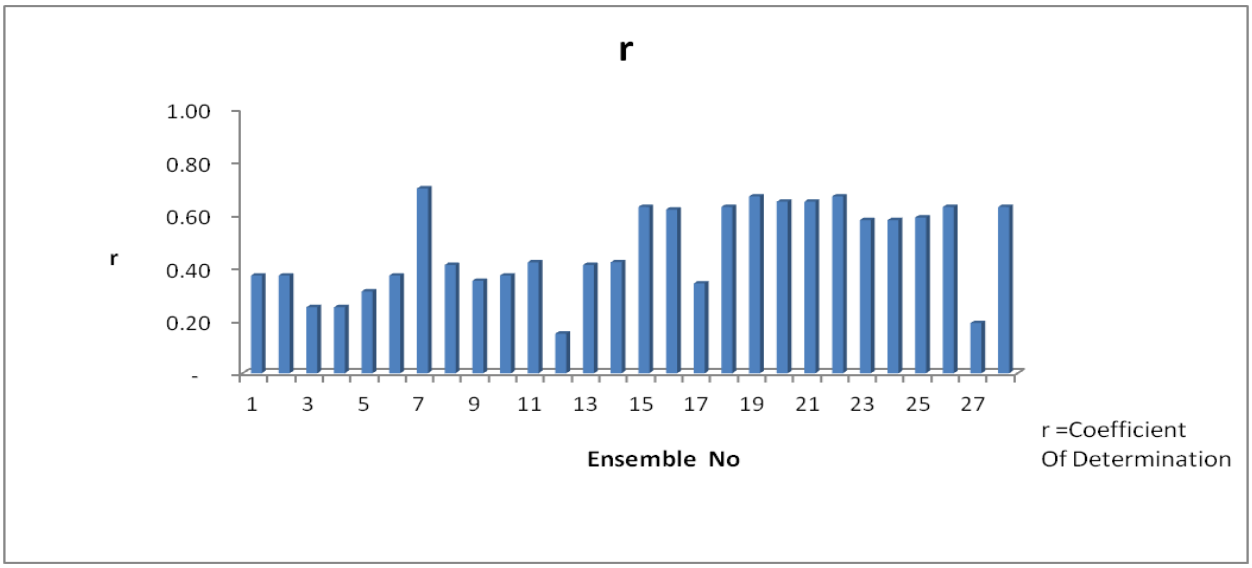

Fig.5.6 r plot with ensembling

Here results of the all the performances measurement parameter like accuracy, rmse, R square, R and graphs between different models are shown in Figure (5.5, 5.6, 5.7, 5.8) respectively.

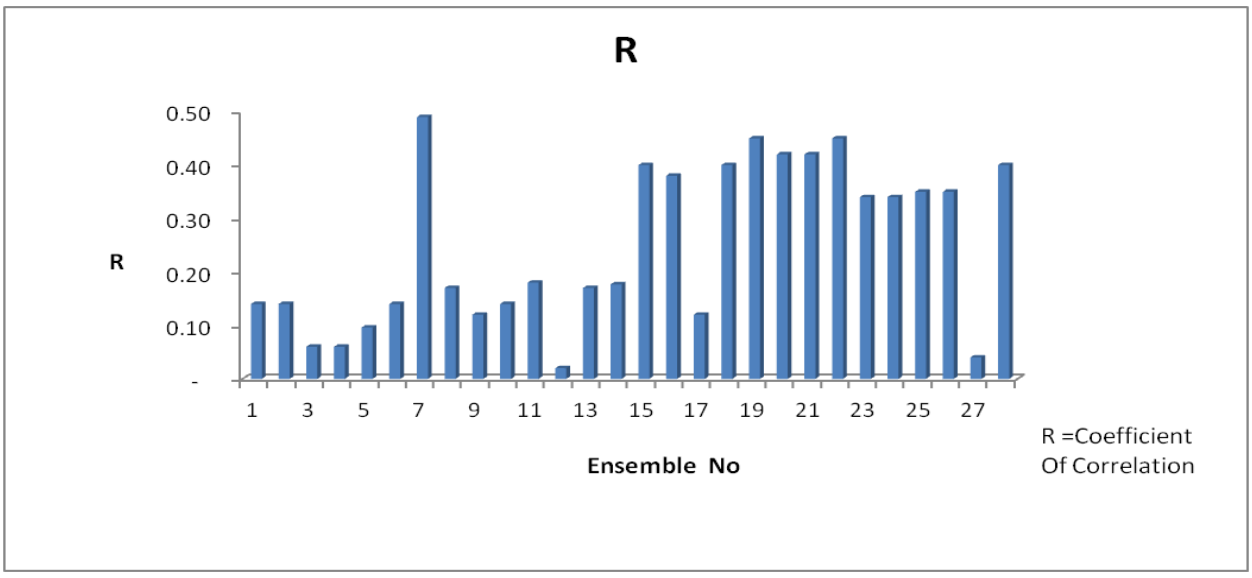

Fig.5.7 R plot with ensembling

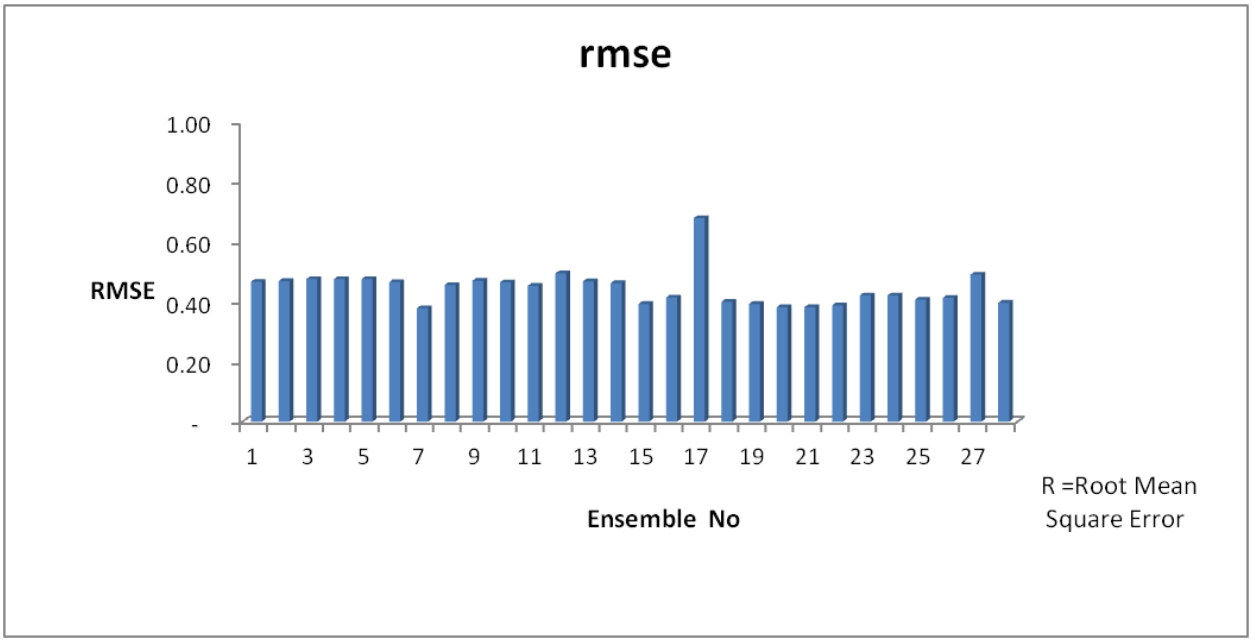

Fig.5.7 RMSE plot with ensembling

Above results show the diversity of different models, here two weak model (Principal Component Regression and Neural Network) having with low accuracy while we ensemble the two model together with the giving best results.

So we apply the same method for second data set to ensure to check diversity factor the results shown below in Table (5.4)

Table 5.4 Separately Results for DS-2

Model Name	r	R	RMSE	Accuracy
RF	0.41	0.33	0.36	89.31
NN	0.12	0.01	0.97	53.7
Enet	0.56	0.26	0.96	80.88
ICR	0.24	0.06	0.52	65.36
PCR	0.17	0.13	0.39	57.92
LR with SS	0.38	0.14	0.48	66.32
LM	0.23	0	0.89	73.32
Cubist	0.45	0.2	0.64	72.3

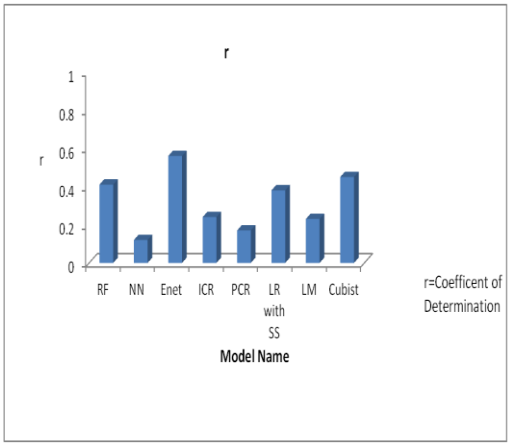

Fig.5.9 r plot for DS2

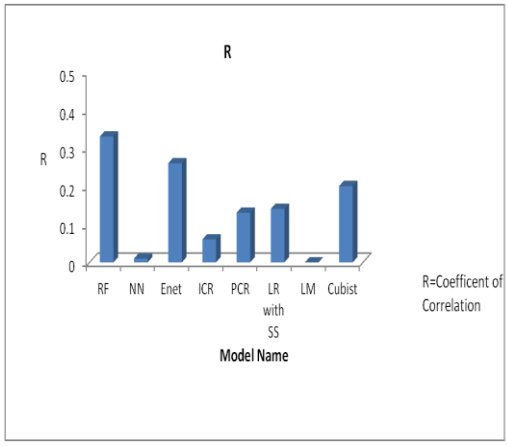

Fig.5.10 R plot for DS2

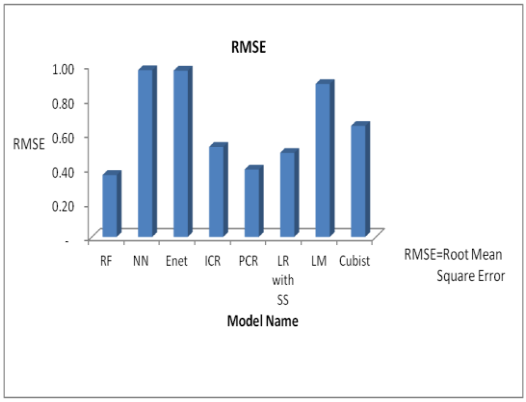

Fig.5.11 RMSE plot for DS2

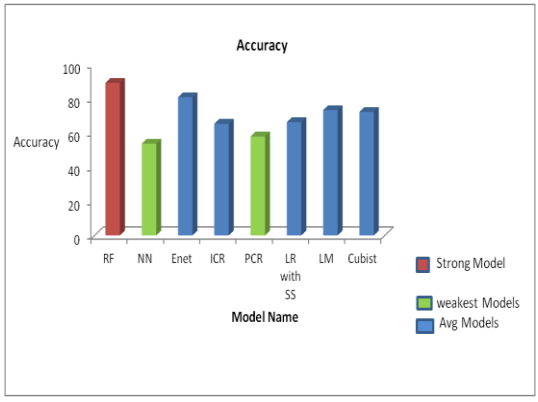

Fig.5.12 Accuracy plot for DS2

Here results of the all the performances measurement parameter like accuracy, rmse, R square, R and graphs between different models are shown in Figure (5.9, 5.10, 5.11, 5.12) respectively.

We have applied the ensembling on Facebook metric performance dataset(DS2). Ensemble technique increase the accuracy of models to prove diversity, results shown in Table 5.5

Table 5.5 Ensemble results for DS2

Model No	**Model Name**	r	R	RMSE	Accuracy
1	NN+PCR	0.15	0.02	0.48	84.00
2	NN+Lr_with_SS	0.17	0.12	0.47	97.21
3	ICR+NN	0.19	0.04	0.47	76.00
4	NN+Enet	0.34	0.41	0.47	97.60
5	NN+LM	0.42	0.18	0.46	91.00
6	PCR + ICR	0.25	0.06	0.49	84.00
7	PCR + ICR	0.25	0.06	0.48	85.00
8	RF+Enet	0.67	0.45	0.39	90.00
9	PCR +Enet	0.37	0.14	0.50	90.62
10	PCR +LR with SS	0.70	0.49	0.38	72.23
11	Enet +icr	0.37	0.14	0.47	81.27
12	ICR+RF	0.62	0.38	0.41	76.65
13	RF+LR_With_SS	0.63	0.40	0.40	87.32
14	PCR + LM	0.31	0.10	0.48	68.32
15	cubist +icr	0.65	0.42	0.38	88.02
16	cubist +LR with SS	0.65	0.42	0.38	66.00
17	NN+Cubist	0.67	0.45	0.39	56.39
18	PCR +Cubist	0.63	0.40	0.40	85.64
19	RF+Cubist	0.58	0.34	0.42	66.32
20	RF+PCR	0.58	0.34	0.42	72.03
21	cubist +lm	0.63	0.40	0.39	80.65
22	icr+lm	0.37	0.14	0.47	85.23
23	LR with SS +icr	0.37	0.14	0.47	84.83
24	RF+LM	0.59	0.35	0.41	84.03
25	RF+NN	0.63	0.35	0.41	81.32
26	Enet+LR_with_ss	0.41	0.17	0.46	80.64
27	LR with SS +LM	0.42	0.18	0.45	76.36
28	LM+Enet	0.35	0.12	0.47	70.25

The graphical representation of the results that is displayed in table are shown below.

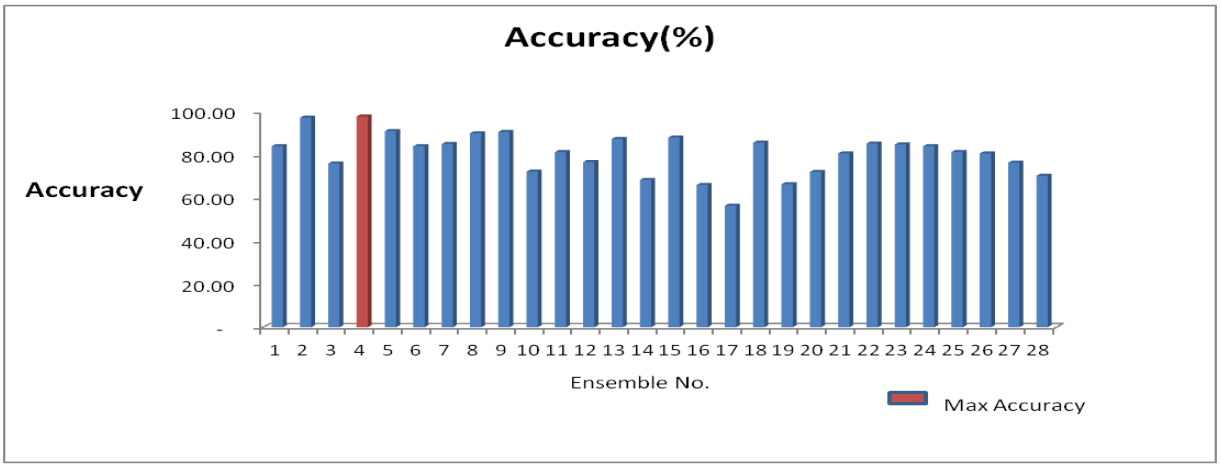

Fig.5.13 Accuracy plot for DS2

From the Figure 5.5 and Figure 5.13 have display the accuracy graph with best results. They shows that those models belong with worst accuracy in separately execution the combination of those models are give best result for both the datasets to prove diversity. The representation with all the measurements parameter are shown in Figure (5.14)

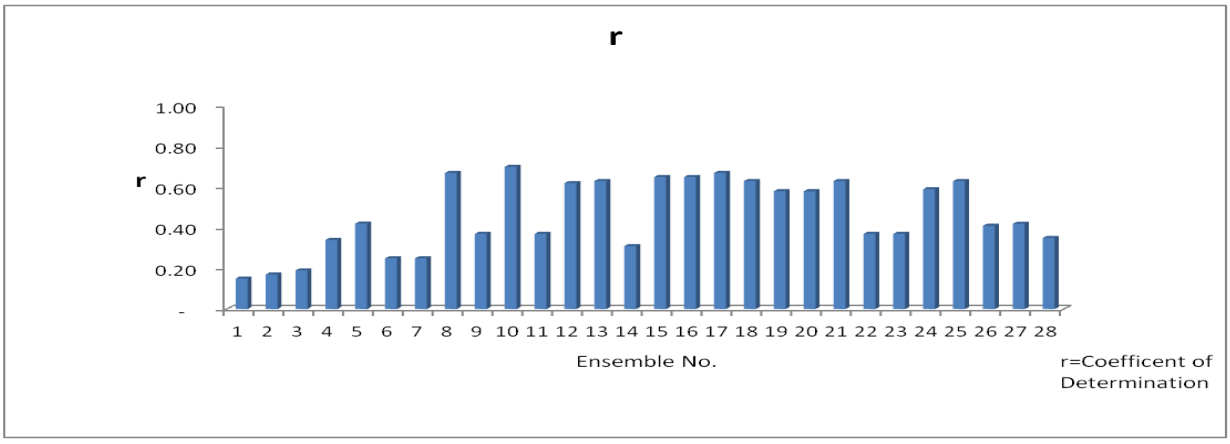

Fig.5.13 r plot for DS2

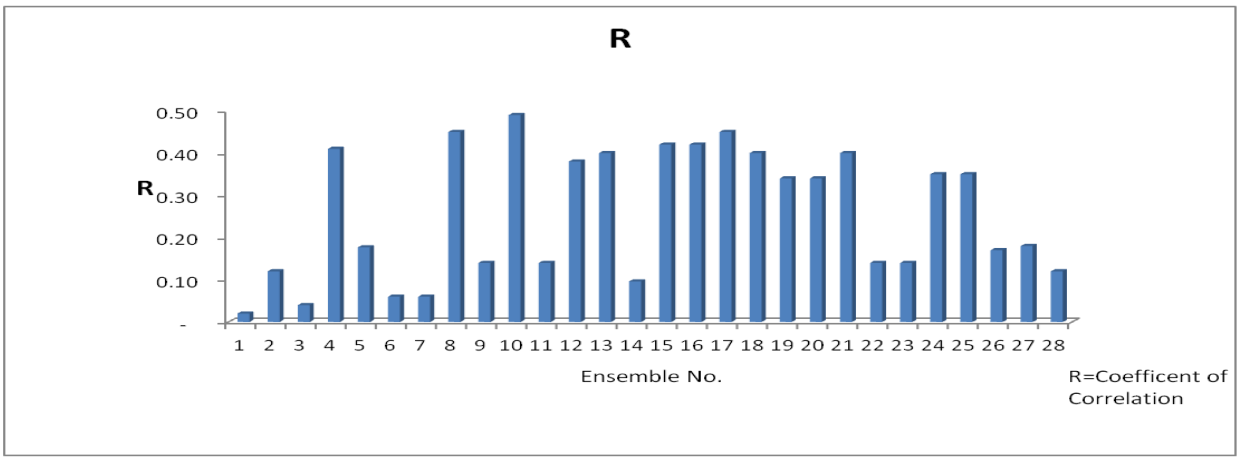

Fig.5.14 R plot for DS2

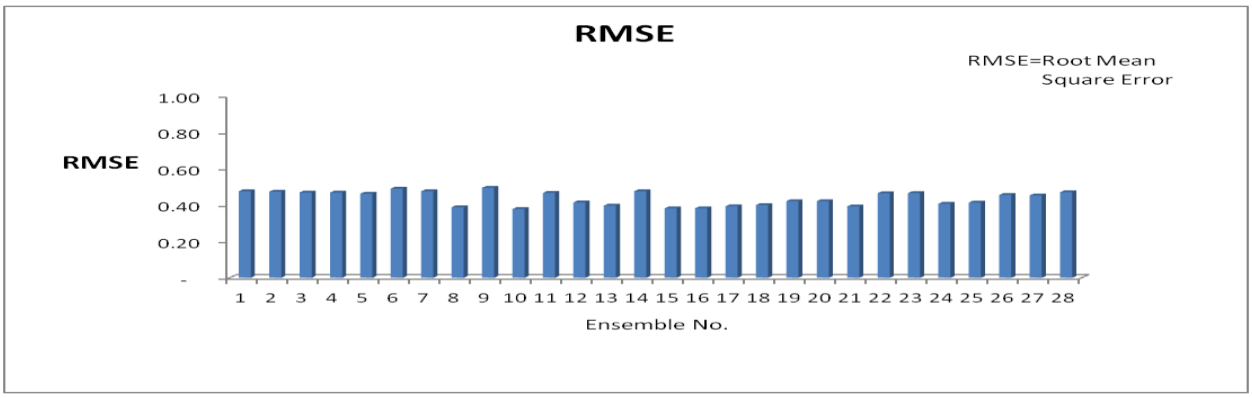

Fig.5.15 RMSE plot for DS

5.3 K-fold cross validation

K-fold cross validation is used to compute the sturdiness of the predictive technique. K-cross validation is a method to calculate the accuracy of a system. For example, take the dataset, D, which is arbitrarily divide into K mutually special 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 total number of the correct classification, which is divided by the number of events occurring in the dataset.

For first dataset based on olfaction prediction select the model with highest accuracy after applied ensembling (i.e. NN and PCR) in which we execute the model ten times for cross check, results are shown in Table(5.6).

Table 5.6 Cross validation results for DS-1

Iteration	r	R	RMSE	Accuracy
I_1	0.15	0.02	0.49	97.8
I_2	0.34	0.13	0.41	97.23
I_3	0.56	0.40	0.40	97.45
I_4	0.19	0.45	0.39	97.34
I_5	0.72	0.42	0.38	97.76
I_6	0.31	0.42	0.38	97.6
I_7	0.62	0.45	0.39	97.08
I_8	0.46	0.34	0.42	97.16
I_9	0.19	0.34	0.42	97.39
I_{10}	0.45	0.35	0.41	97.33

For the dataset based on olfaction prediction select the model with highest accuracy after applied ensembling (i.e. NN and PCR) in which we execute the model ten times for cross check and the robustness of ensemble as discussed in graphical representation shown in Figure (5.16) .

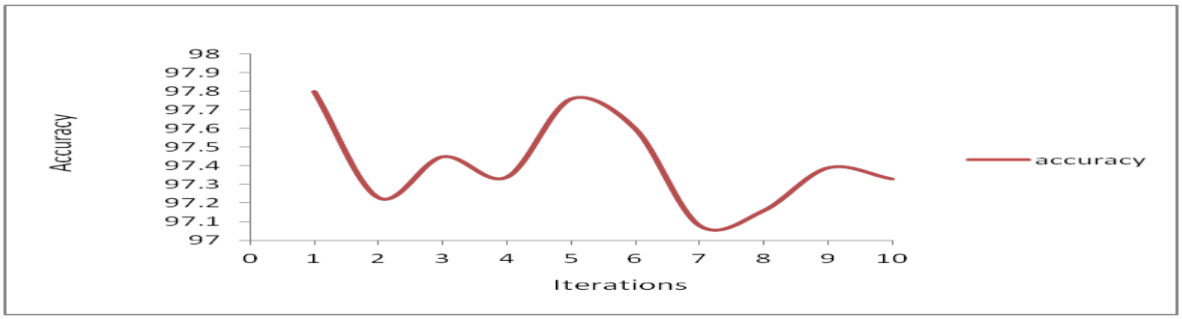

Fig.5.16 Cross validation on NN and PCR

Chapter 6: Conclusion and Future Scope

6.1 Conclusion

Improving classification accuracy in the ensemble is a major problem in real world datasets. Numbers of ensemble techniques like bagging, boosting, and stacking, are presented to produce diversity in the ensembles. However, these techniques suffer from data loss. When in bagging technique if increase the size of data sets, then we can't improve the model performance, but we can decrease the variance. Surmount the diversity ensemble problem in this thesis; a framework is proposed, called Diversity Ensemble (DEF) framework. The classifier's algorithms are given the best results in terms of accuracy after getting the single model's results we divide all them into two categories on the basis of their accuracy first one is weak bucket model and another one is strong bucket models. After that, we have applied classification algorithms on balanced datasets and with the help of various R packages and libraries, we select the important features. Further, we applied classification algorithms on selected features and perform ensembling on all the classifiers. When a single model for certain dataset is used for building ensemble in a group of two then, it is observed that the two models which belongs to the weak performing category are giving the best results as compare to the model which are belonging to strong model category. It can be improve significantly by building the ensemble of weak classifiers. Finally, it is concluded that proposed framework gives better performance results, and successful in getting odor prediction for the datasets.

6.2 Future Scope

For future, we would experiment the impact of diversity in building ensemble for high dimensional and complex big datasets.

One more extension to our work could be using prediction ranking for choosing classifiers for dynamic ensemble construction. Each classifier's predictions could be ordered by the associated confidence value. After the classifier predicts a new data

instance with some confidence, we could find the ranking of this new instance and choose classifiers based on these ranking values.

In future we can verify our ideas with more sets of data; perform experiments with more ensemble methods; integrate all the experiment environments into one platform.

References

[1] R. E. Schapire, “The strength of weak learnability”, Machine Learning, vol. 5, no. 2, pp. 197-227, 1990.

[2] Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy", Machine Learning vol. 51, no. 2, pp. 181-207, 2003.

[3] L. Lam, “Classifier Combinations: Implementations and Theoretical Issues”, Multiple Classifier Systems Lecture Notes in Computer Science, pp. 77–86, 2000.

[4] Bauer, Eric, and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants", Machine Learning vol. 36 no. 1 pp. 105-139,1999.

[5] P. C. A. Cunningham and J. Carney, “Diversity versus Quality in Classification Ensembles Based on Feature Selection”, Machine Learning: ECML 2000 Lecture Notes in Computer Science, pp. 109–116, 2000.

[6] Whitley, L. Darrell, J. Ross Beveridge, Cesar Guerra-Salcedo, and Christopher R. Graves. "Messy Genetic Algorithms for Subset Feature Selection", In ICGA, pp. 568-575. 1997.

[7] D. Brzezinski and J. Stefanowski, “Ensemble Diversity in Evolving Data Streams”, Discovery Science Lecture Notes in Computer Science, pp. 229–244, 2016.

[8] S. Wang and X. Yao, “Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 206–219, 2013.

[9] L. I. Kuncheva, “Combining pattern classifiers: methods and algorithms”, Hoboken, NJ: Wiley-Interscience, 2004.

[10] K. Tumer and J. Ghosh, “Error Correlation and Error Reduction in Ensemble Classifiers” Connection Science, vol. 8, no. 3-4, pp. 385–404, 1996.

[11] G. Zenobi and P. Cunningham, “Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error”, Machine Learning: ECML 2001 Lecture Notes in Computer Science, pp. 576–587, 2001.

[12] Krogh, Anders, and Jesper Vedelsby. "Neural network ensembles, cross validation, and active learning." Advances in neural information processing systems vol. 7, pp. 231-238, 1995.

[13] P. Melville and R. J. Mooney,"Creating diversity in ensembles using artificial data," Information Fusion, vol. 6, no. 1, pp. 99–111, 2005.

[14] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. Kegelmeyer, "Ensemble diversity measures and their application to thinning," Information Fusion, vol. 6, no. 1, pp. 49–62, 2005.

[15] P. C. A. Cunningham and G. Zenobi, "Case Representation Issues for Case-Based Reasoning from Ensemble Research," Case-Based Reasoning Research and Development Lecture Notes in Computer Science, pp. 146–157, 2001.

[16] Huang, Gao, Shiji Song, Jatinder ND Gupta, and Cheng Wu. "Semi-supervised and unsupervised extreme learning machines" IEEE Transactions on Cybernetics vol. 44, no. 12, pp. 2405-2417, 2014.

[17] Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al.., eds.; 2006)[book reviews]" *IEEE Transactions on Neural Networks* vol. 20, no. 3, pp. 542-542, 2009.

[18] "IOSPress," *IOS Press*. [Online]. Available: http://www.iospress.nl/catalogue/Bookseries/frontiers in artificial intelligence and applications [Accessed: 14-May-2017].

[19] C. E. Rasmussen, "Gaussian Processes in Machine Learning," *Advanced Lectures on Machine Learning Lecture Notes in Computer Science*, pp. 63–71, 2004.

[20] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, 2015.

[21] Kubat, Miroslav, Robert C. Holte, and Stan Matwin. "Machine Learning for the detection of oil spills in satellite radar images." *Machine Learning*, vol. 30, no. 2-3, pp. 195-215, 1998.

[22] Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan."An introduction to MCMC for Machine Learning." *Machine Learning*, vol.50, no. 1, pp. 5-43, 2003.

[23] Hong, CharmGil. "Gaussian Processes in Machine Learning." (2011).

[24] Mitchell, Tom Michael. “The discipline of Machine Learning” Carnegie Mellon University, School of Computer Science, Machine Learning Department, Vol. 9, 2006

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

[26] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” *Journal of Big Data*, vol. 2, no. 1, 2015.

[27] E., “Six Novel Machine Learning Applications,” *Forbes*, 06-Jan-2014. [Online]. Available: https://www.forbes.com/sites/85broads/2014/01/06/six-novel-machine-learning-applications/#26c2ac911060. [Accessed: 14-May-2017].

[28] Dietterich, Thomas G, "Ensemble methods in Machine Learning.", In *International workshop on multiple classifier systems*, Springer Berlin Heidelberg, pp. 1-15, 2000.

[29] R. E. Schapire, “The strength of weak learnability”, *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[30] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: a new explanation for the effectiveness of voting methods,” *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[31] Hansen, Lars Kai, and Peter Salamon, "Neural network ensembles", *IEEE transactions on pattern analysis and machine intelligence*, vol.12, no. 10, pp. 993-1001, 1990.

[32] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[33] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[34] Freund, Yoav, and Robert E. Schapire, "Experiments with a new boosting algorithm" In *icml*, vol. 96, pp. 148-156. 1996.

[35] Keller, Andreas, Richard C. Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D. Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, Celine Vens, Leander Schietgat, Kurt De Grave, Raquel Norel, Gustavo Stolovitzky, Guillermo Cecchi, Leslie B. Vosshall, and Pablo Meyer. "Reverse-engineering human olfactory perception from chemical features of odor molecules." (2016): n. pag. Web.

[36] Moro, Sérgio, Paulo Rita, and Bernardo Vala. "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach." *Journal of Business Research* 69.9 (2016): 3341-351. Web.

[37] Wolpert DH, "Stacked generalization", Neural Netw ,vol. 5, no. 2, pp. 241–259, 1992

[38] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. Kegelmeyer, "Ensemble diversity measures and their application to thinning," *Information Fusion*, vol. 6, no. 1, pp. 49–62, 2005.

[39] "Sage Synapse : Contribute to the Cure." *Sage Synapse : Contribute to the Cure*. N.p., n.d. Web. 13 March 2017.

[40] *UCI Machine Learning Repository*. N.p., n.d. Web. 23 March 2017.

[41] **R**ana, Prashant Singh, Harish Sharma, Mahua Bhattacharya, and Anupam Shukla. "Quality assessment of modeled protein structure using physicochemical properties." *Journal of Bioinformatics and Computational Biology* 13.02 (2015): 1550005. Web.

[42] Leo Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[43] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. The measurement of inter-rater agreement. Statistical Methods for Rates and Proportions, 2:212–236, 1981

[44] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[45] E Ke Tang, Ponnuthurai N Suganthan, and Xin Yao. An analysis of diversity measures. Machine Learning, 65(1):247–271, 2006.

[46] Ludmila I Kuncheva, James C Bezdek, and Robert PW Duin. Decision templates for multiple classifier fusion: An experimental comparison. Pattern Recognition, 34(2):299– 314, 2001.

[47] Derek Partridge and W Krzanowski. Software diversity: Practical statistics for its measurement and exploitation. Information and Software Technology, 39(10):707–717, 1997.

[48] Yi Zhang, Samuel Burer, and W Nick Street. Ensemble pruning via semi-definite programming. The Journal of Machine Learning Research, 7:1315–1338, 2006.

[49] Ron Kohavi and David H Wolpert. Bias plus variance decomposition for zero-one loss functions. In Machine Learning: Proceedings of the Thirteenth International, pages 275–283, 1996.

[50] Dragos D Margineantu and Thomas G Dietterich. Pruning adaptive boosting. In International Workshop on Machine Learning, volume 97, pages 211–218, 1997

[51] Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In International Workshop on Machine Learning, volume 96, pages 148–156. Morgan Kaufmann, 1996.

[52] Lars Kai Hansen and Peter Salamon. Neural network ensembles. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 12(10):993–1001, 1990.

[53] Prem Melville and Raymond J Mooney. Creating diversity in ensembles using artificial data. Information Fusion, 6(1):99–111, 2005.

[54] Marcelo N Kapp, Robert Sabourin, and Patrick Maupin. An empirical study on diversity measures and margin theory for ensembles of classifiers. In Tenth International Conference on Information Fusion, pages 1–8. IEEE, 2007.

[55] Kuo-Wei Hsu and Jaideep Srivastava. Relationship between diversity and correlation in multi-classifier systems. In Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, editors, Advances in Knowledge Discovery and Data Mining, volume 6119 of Lecture Notes in Computer Science, pages 500–506. Springer Berlin Heidelberg, 2010.

[56] Gavin Brown and Ludmila I. Kuncheva. "good" and "bad" diversity in majority vote ensembles. In Neamat Gayar, Josef Kittler, and Fabio Roli, editors, Multiple Classifier Systems, volume 5997 of Lecture Notes in Computer Science, pages 124–133. Springer Berlin Heidelberg, 2010.

[57] Gavin Brown. An information theoretic perspective on multiple classifier systems. In Multiple Classifier Systems, pages 344–353. Springer, 2009.

List of Publications & Video Link

[1] Ashish Gour and Seema Bawa, **"*Olfaction prediction of Chemicals using Ensemble Machine Learning"***, in *IEEE International Conference on Information Communication, Instrumentation & Control*" (IEEE) (ICICIC - 2017) [Communicated].

[2] https://www.youtube.com/watch?v=RFEL-8CrWiA

ASHISH_801531005

ORIGINALITY REPORT

%9	%5	%6	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Sašo Džeroski. "Machine Learning, Ensemble Methods in", Computational Complexity, 2012 Publication	%1
2	www.authorea.com Internet Source	<%1
3	www.cs.utexas.edu Internet Source	<%1
4	machinelearningmastery.com Internet Source	<%1
5	Rana, Prashant Singh, Harish Sharma, Mahua Bhattacharya, and Anupam Shukla. "Quality assessment of modeled protein structure using physicochemical properties", Journal of Bioinformatics and Computational Biology, 2014. Publication	<%1
6	www.cs.bme.hu Internet Source	<%1
7	Carl Edward Rasmussen. "Gaussian Processes	