THE ANALYSIS AND IMPACT OF CO2 WORLDWIDE

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

JAGTAR SINGH

11501440

Supervisor

ASST. PROF MD. ATAULLAH

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

January – May 2017

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0260

Supervisor Name : Md. Ataullah **UID :** 16915 **Designation :** Assistant Professor

Qualification : M.Tech **Research Experience :** 5 Years

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Jagtar Singh	11501440	2015	K1518	8054502648

SPECIALIZATION AREA : Networking and Security **Supervisor Signature:** Md. Ataullah 16915 22/06/2017

PROPOSED TOPIC : The analysis and impact of CO2 emissions Worldwide.

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.33
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	8.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.67
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.00
5	Social Applicability: Project work intends to solve a practical problem.	8.33
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.33

PAC Committee Members		
PAC Member 1 Name: Prateek Agrawal	UID: 13714	Recommended (Y/N): Yes
PAC Member 2 Name: Pushpendra Kumar Pateriya	UID: 14623	Recommended (Y/N): Yes
PAC Member 3 Name: Deepak Prashar	UID: 13897	Recommended (Y/N): NA
PAC Member 4 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): NA
PAC Member 5 Name: Anupinder Singh	UID: 19385	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: **The analysis and impact of CO2 emissions Worldwide.**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 03 May 2017

ABSTRACT

Carbon plays an essential role in the environment for climate change. The presence and absence of carbon directly affects all living beings. Trees inhale carbon for giving us oxygen. The environmental study of carbon is a major concern these days. Carbon cycle is essential to life on earth but it is recognized as one of the most dominating gas among greenhouse gases (GHG). Each country emits varying amount of carbon dioxide each year from different sources including burning of oil, coal and gas as well as deforestation. Each year a rank is also given to each country based on statistical measures of education, poverty, life expectancy and income levels known as human development index (HDI). The focus of this study is to predict the level of development for each country based on human development index and carbon emission by each year and categorize them to three different classes using algorithms like Naïve Bayes, K- nearest neighbor (KNN), Decision tree, Random forest and J48. By analysing the results of these algorithms on year wise data set, it is concluded that tree based algorithms are more accurately able to classify the data.

Keywords: *Classification, Human Development Index (HDI), Carbon Dioxide, Naive Bayesian Classifier, Decision Tree, K-Nearest Neighbor (KNN), Random Forest, J48.*

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "THE ANALYSIS AND IMPACT OF CO2 WORLDWIDE" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. "Md. Ataullah" I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

JAGTAR SINGH

11501440

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M. Tech Dissertation entitled "**THE ANALYSIS AND IMPACT OF CO2 WORLDWIDE**", submitted by **Jagtar Singh** at **Lovely Professional University; Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Mr. Md.Ataullah

Date:

Counter Signed by:

1) **HoD's Signature:** ________________

HoD Name: ____________________

Date: ____________________

2) **Neutral Examiners:**

(i) Examiner 1

Signature: _______________

Name: ___________________

Date: ____________________

(ii) Examiner 2

Signature: _______________

Name: ___________________

Date: ____________________

ACKNOWLEDGEMENT

I wish to acknowledge my sincere indebtedness and a deep sense of gratitude to Mr. Md. Ataullah, Assistant professor of Database System, School of Computer Science and Engineering, LPU Jalandhar (Punjab) for his generous guidance, help, useful suggestions and making my work a pleasant experience. I shall always visualize his encouraging gesture, constant inspiration, and constructive and indefatigable zest to learn increasingly.

I express gratitude Head of School and other faculty members of Computer Science and Engineering Department, LPU for their intellectual support throughout the course of this work.

The words are not adequate to express my gratitude to my parents and entire family for their patience, support and active assistance in many ways. Finally, I indebted to all my colleagues, dears & nears and all whosoever has contributed in this dissertation work directly or indirectly.

Jagtar Singh

11501440

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 CARBON DIOXIDE – INTRODUCTION

The Carbon Cycle is an arrangement that implies the productive relationship of carbon by means of air and living being from water and land. To life on earth, this cycle of carbon is very basic. Carbon Dioxide is perceived as the best overwhelming gas amongst greenhouse gases (GHG). The amount of carbon in air is only 0.04 [12]. The general study pattern of recognizing the emission of carbon level is expanded since years of 2000 to 2014. This survey implies that emission of carbon is a major area of concern for tending the change of environment. Since population development and expanding population influences the climate, so analysis of ecological change is fundamental area for research in future. Not just growing countries like China, India and Brazil are focusing on this examination yet the developed countries, for instance, USA, Germany, UK, Canada and so forward and sum of 133 countries present to this exploration [22]. The following figure shows that how much carbon emission is done by different production sectors like energy supply 26, Transport 13, Deforestation 17, and Agriculture 14 etc.

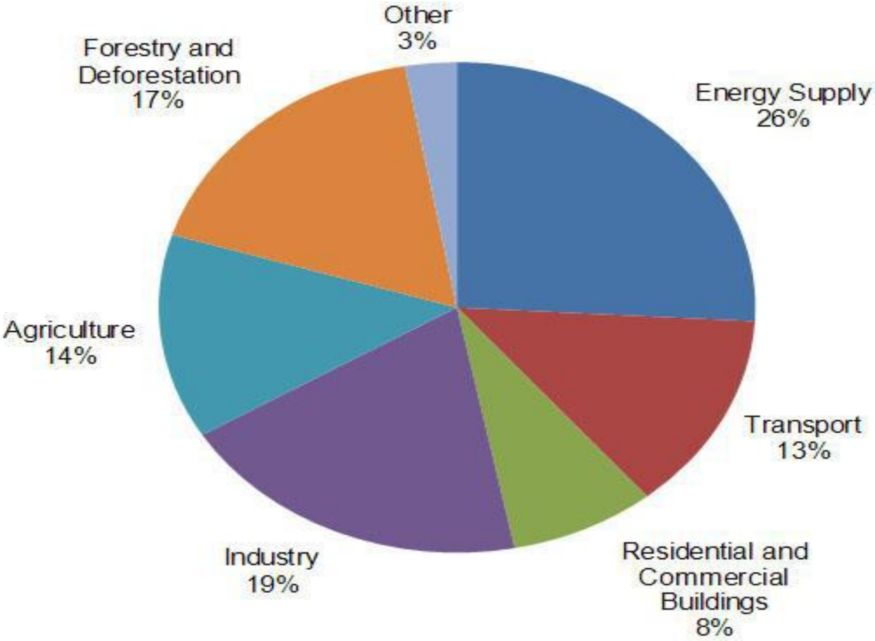


Figure 1.1: Percentage of worldwide carbon emission from numerous sources

1.2 OVERVIEW OF HUMAN DEVELOPMENT INDEX (HDI)

Human Development Index (HDI) is an effort to imitate humanoid prosperity and improvement encounters of people as individuals, as well as people from a gathering or group, national or country past per capita salary or all variations of it. This is very common movement of another improvement discourse called 'human advancement worldview\ paradigm' exhibited by UNDP in 1990 in its first Human Development Report. The notions of human capacities as well as working, this report expel the pay as the exclusive measuring stick of improvement, place individuals and its prosperity by the point of convergence of advancement technique [1].

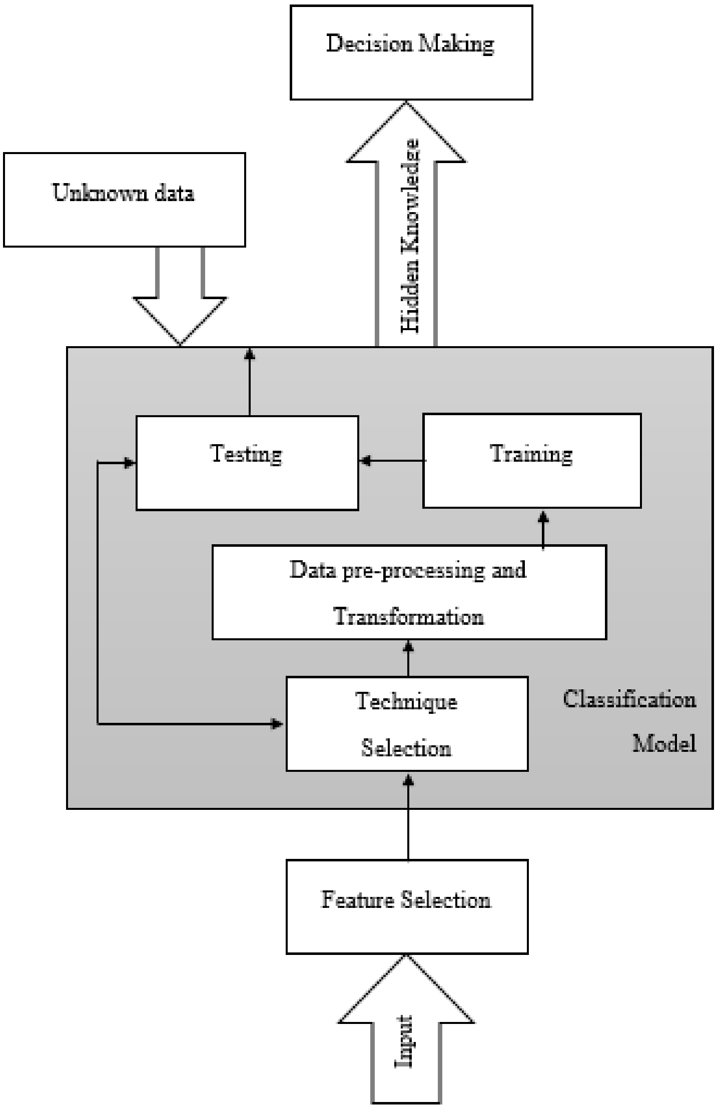
This development report i.e. HDR 1990 characterized individuals growth as a "procedure of broadening individual's decisions through extending abilities of people as well as functioning [1]. The most accusing of these widespread decisions are transmit on with an extended and vigorous life that is to be instructed to approach assets required for an average way of life. On the off chance that these basic decisions are not accessible, numerous different open doors stay blocked off. Dismissing revenue as the one and only quantity of prosperity and development. HDR 1991 perspectives that revenue is unique part of these selections and a critical one yet it is not entirety of individuals life and do not require boundless salary to have decent existence. In this way, along with the idea of development of individual, the UNDP additionally presented following the HDI by joining there fundamental measurements of lifespan.

- Knowledge-*the ability to attain knowledge and be educated.*
- Longevity -*the capability to lead an extensive and vigorous life.*
- Income - *the ability towards appreciate a decent living standard and have a socially significant life* [1].

The index is utilized for presenting the human development world summary of nations and accordingly in positioning as well as arranging these in light of scores of HDI. From that point forward, the HDI has dragged in incredible deal of attraction from policy makers, improvement masterminds, activists and its backers [1]. In response to the enhanced perspectives and conclusions, the technique of the HDI has experienced a few changes, occasionally, recipes, goalposts and others. The reason is to highlight variations that have been presented HDI and the development that has occurred in it.

1.3 INTRODUCTION TO DATA MINING

Data mining is the way toward separating valuable information. It is the process of finding concealed patterns and information from the existing data[2]. In data mining, one needs to essentially focus on purifying the data to make it achievable for additional processing. The process of purifying the data is additionally called as noise elimination or noise lessening or feature elimination. Likewise, Data mining is the procedure of utilizing collection of analysis of data instruments from information to find patterns in addition to connections that might be utilized to create legitimate expectations. Additionally, characterize this as an iterative and intelligent procedure of finding substantial, unique, valuable, and understandable patterns in gigantic database.



.

Figure 1.2: The General Data Mining Process Model

1.3.1 Techniques of Data Mining

To evaluate huge amount of data, data mining came into picture and is also called as KDD process. To finish this procedure different strategies grew so far are clarified in this segment. KDD is the general procedure, which is appeared in figure 3.

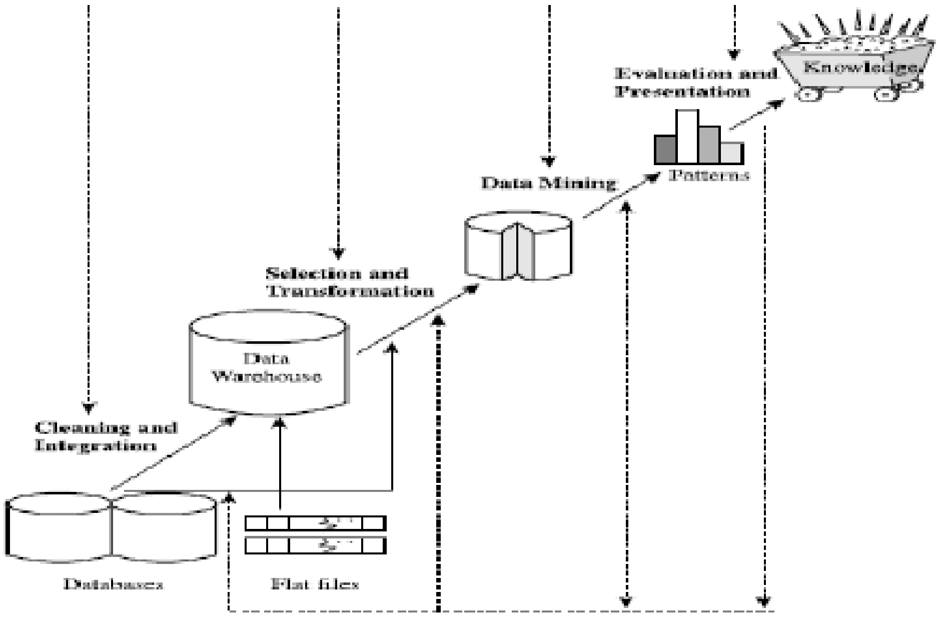


Figure 1.3: Knowledge Discovery Process [2]

In KDD, the fundamental and essential step is data mining. KDD will transform the low-level data into abnormal state data. Data mining is the documented in which valuable result that is being predicted from vast database. It utilizes effectively fabricated apparatuses to get out the helpful hidden patterns, trends and expectation of future can be obtained using the procedures [3]. Data mining includes model to find patterns, which comprises of different components.

The following are number of data mining techniques:

- Clustering
- Classification
- Regression
- Association rules

i. Clustering

Clustering is defined as recognizable credentials of comparative modules of objects. By utilizing this, methods identify the dense and inadequate areas in space of object and can find general dispersion pattern and connections amongst features of data. It is unsupervised technique of classification or it is otherwise called exploratory data analysis in which there is no arrangement of labeled data. The fundamental aim of clustering procedure is to isolate the unlabeled data set into limited and discrete arrangement of characteristic and hidden data structures[2]. There is no arrangement of providing precise representation of imperceptibly samples that are created from by same probability distribution. Classification approach can also be utilized for powerful methods for recognizing collections or sessions of entity however turns it out to be expensive so it can be utilized as preprocessing method for property subdivision assortment and organization [2]. For instance, to shape collection of clients in light of acquiring patterns, to classifications qualities with comparative functionality.

ii. Classification

This is the most usually associated technique of data mining, which utilizes an arrangement of characterized cases that develop a prototype, which can group the number of records at huge amount. The applications like detection of fraud and credit risk are particularly appropriate to this kind of consideration. This methodology frequently uses decision tree or neural network-based classification techniques [3]. The process of classification of data includes learning and classification. The training data were examined by algorithm of classification in learning. The test data were used to evaluate the exactness of rules in classification. If accuracy is acceptable, the guidelines can be associated with the new tuples of data. This would incorporate records of both fake and legitimate actions decided on a record-by-record premise [3]. The algorithm of classifier training uses the pre-characterized cases that decide the arrangement of required parameters for appropriate segregation. This technique at that point scrambles the factors into a model that is a classifier.

iii. Regression

The analysis of this can utilized just before display the connection amongst at least one autonomous factors and dependent variables. This technique can be adapted from predication. Autonomous factors are characteristics definitely recognized and reaction factors are that need to forecast in data mining. For example, trades dimensions, store costs, in addition to item disappointment charges that all are exceptionally hard to expect in light of the fact that they may rely on difficult collaborations of numerous forecaster factors. Consequently more composite methods e.g. neural network , logistic regression, decision trees, might be important to estimate forthcoming esteems[3]. For both regression and classification, a similar model types can regularly be utilized. For instance, the Classification and Regression Trees, decision tree procedure can be utilized to assemble trees of classification to group straight out reaction factors and regression trees to figure persistent reaction factors. Neural networks also make jointly classification and regression prototypes.

iv. Association rules

This is typically used to determine recurrent set of item discoveries surrounded by extensive set of data. It is also called correlation method. This kind of discovering encourages organizations to settle on specific results, for example design of directory, cross advertising and analysis client shopping behavior [2]. The Association Rule algorithm should have the capacity to produce procedures with less certainty value. On the other hand, the quantity of conceivable rules of association for a specified dataset is for the most part expansive and a high extent of the rules are more often than not of little assuming any esteem.

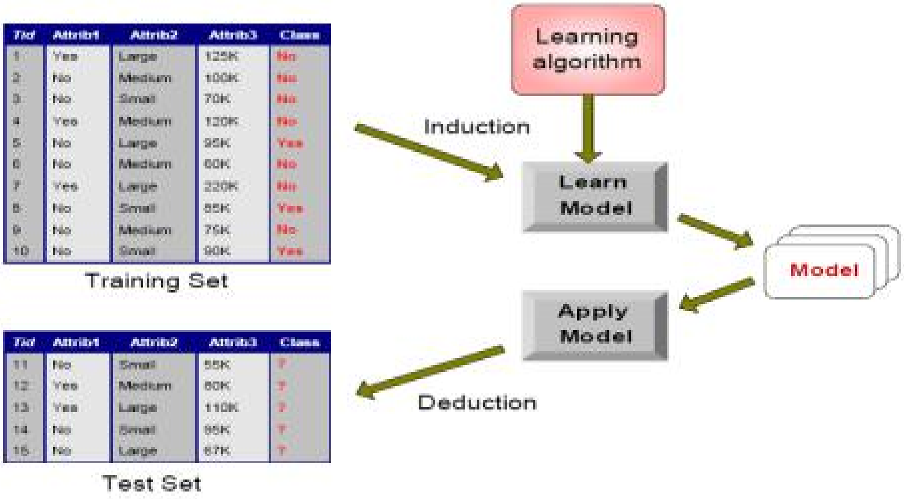
1.4 OVERVIEW OF CLASSIFICATION

Classification is a function of data mining that allocates items in a collection to target categories or classes. For each case in the data, the goal of this technique is to predict the target class. Further, it is the method of the data mining that is fundamentally used to evaluate a dataset that is given; every instance of this is taken as well as exchanges this occurrence to a particular class with the end goal that error of classification will be scarcest. Furthermore, it is a procedure which comprises of predicting a particular outcome in presence of given input information [4]. Also to classify each and every detail into one of previously recognized set of classes or groups in an arrangement of data. There is a necessity of one training set, which comprises of an arrangement of properties, and the individual results called the goal attribute to determine the predefined yield. The procedure attempts to determine associations amongst the characteristics that would make it probable to visualize the result. At that point, an unknown dataset is given to the algorithm, which comprises of same arrangement of attribute with the exception of the characteristic, which is unknown. This procedure examines information data of input as well as produces the particular result. Classification has two types: Supervised Classification and Unsupervised Classification [5]. Supervised classification is one of the primary techniques to extricate information from databases where set of training examples are known already and in unsupervised classification training examples are not known previously and the results depends on the product analysis of an image without the client providing sample classes.

Actually, Classification is a twofold procedure, which comprises two stages.

- One is Testing phase where testing of classifier is done to analyze its performance using various samples of the test set [5]. Prediction accuracy is a rule to assess the performance of classifier. Classification accuracy depicts the rate of instances, which are correctly grouped.

- The other is Training phase where with the help of classifier algorithm, Training dataset trains the classifier and discovers connections amongst the predicators values and the target values.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Figure 1.4: illustrating classification task [4].

1.5 SELECTED CLASSIFICATION PROCEDURES

This segment talks about the ideas and standards of the methods recognized in doing classification. The following classification techniques are used to analyze and to find impact of carbon dioxide using global dataset.

1.5.1 Decision Tree (DT)

DT models are usually applied as a part of data mining. It is utilized to observe data and originate the tree and its rules, which will be used to create forecasts [5]. It is prominent to be a powerful method of classification in a few areas. This is a representation method, which is used to arrange of rules that prompt to a value and class. The forecast of this could be to foresee every single absolute assets when instances are to be put in classifications or classes.

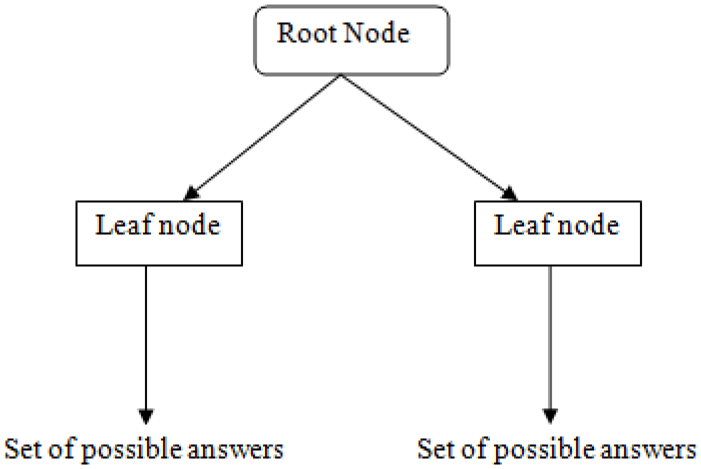


Figure 1.5: Decision tree induction[4]

The procedure could likewise be utilized as a part of the forecast of proceeds with variable relapse trees that total values are necessary. It is created through procedures of iteration of splitting information into discrete classes. The objective is to amplify the separation amongst bunches at each split. It depends on the procedure used to execute it that how to perform splitting. It is possible to build the huge number of DTs as would be prudent from a given properties of set of data. Finding the ideal tree is computationally tremendous when the space of search is extensive, then these trees are more correct than others are, [5]. Productive algorithm has been created to establish to incite a sensible precision inside sensible measure of time. For instance, these procedures are the Hunt's algorithm, which structures the bases of existing algorithms of decision tree induction that includes C4.5. These techniques are more frequently used than strategy of greedy algorithm in searching the characteristics space and use for the information portioning. This point is plot by how Hunt calculation capacities from a strange state viewpoint. A tree is created in a way of recursion by isolating the dataset into dynamic subsets [5] [6].

Supposing Ts is the arrangement of records of instances of training which are related with node s and y = {y1, y2…yc} represents labels of class. The Hunts procedure recursively describes the following:

- If every one of the instances in Ts have a similar class Y1 then s is the node i.e. leaf node marked as Ys
- If Ts comprises instances that have a place with other than a class, a property test situation is chosen to segment the occurrence into a smaller subset and the instances in Ts are appropriated to the children in view of the result. This is connected to every child node.

The yield of a decision tree is transparent that makes it understandable and simple for clients or non-specialized people to recognized.[5] Decision tree strategies are identified to have versatility and productivity issues, for example, significant reduction in execution and poor utilization of accessible framework assets.

1.5.2 Random Forest

Random Forests is a technique by which one can compute precision rate in better way it is a collection of tree indicators called forest [5]. Random Forests develops numerous classification trees. The random trees classifier takes the input feature vector, characterizes it with each tree in the forest, and yields the class label that received the dominant part of "votes". Position the information vector down each of the trees in the forest to group a new object from an input vectors. Each tree performs a classification, and called tree as "votes" for that class. For performing out no pruning choose a test that is based on random elements at every node. Random forest creates random woods by stowing troupes of random trees. The forest picks the classification having the most votes over each tree in the forest. Each tree is developed as follows: If the quantity of cases in the training set is N, test N cases at random however with substitution, from the first information. This same t part on this m is used to split the node. The estimation of m is held consistent during the development of forest.

1.5.3 Decision tree algorithm J48:

This approach is extreme valuable in classification issue. A tree is developed for demonstration of classification procedure. It is applied to database to every tuple which results classification of data once the tree is constructed. J48 is an extension of ID3. It makes a binary tree. It is a basic C4.5 decision tree [7]. The J48 classifier algorithm works, with a specific goal to classify a new item, it initially needs to make a decision tree in view of the attribute values of the accessible training data [13] [17]. Thus, at whatever point it encounters an arrangement of items training set it distinguishes the attribute that discriminates the different occurrences most clearly. The component that can disclose to tell us most about the data instances so that can arrange them the best is said to have the highest information gain [14]. J48 overlooks the missing values in which incentive for that item can be predicted based on that is known about the attribute values for alternate records while building a tree,. The essential notion is to isolate the range from data based on the property values for that thing that are found in the training sample. The classification of J48 permits by means of either decision trees or rules produced from them.

1.5.4 Naïve Bayes (NB)

It is totally based on the Bayesian theorem. This strategy of classification analyses the association amongst each property and the class for each case to decide a restrictive probability for the associations amongst the characteristic values and the class. The rule behind Naïve bayes is a basic procedure. The probability of each class is processed by numbering that how frequently it happens in the training dataset [12]. This is known as the "prior probability" P(C=c). The algorithm additionally makes sense of the probability for the event x given c having the supposition that the characteristics are autonomous. This probability transforms into the result of the probabilities of each characteristic.

Bayes theorem offers a technique for processing the posterior probability $P(c|x)$ from P(c), P(x), and $P(x|c)$ [12]. This considers that the impact of the value of a predicator (x) on a given as:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c) \quad (1)$$

- $P(c|x)$ is the posterior probability of class (target) given predictor (attribute) of class.
- P(c) is called the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class.
- Class (c) the values of other predictors i.e. independent [12].

It is well-known to be tremendously effective technique and it acquires in a linear fractions using association rule. On the other hand, when features are redundant and not ordinarily circulated this leads to influence the accuracy of prediction. Little work has been done as far as to solve data mining issues with the use of Naïve Bayes. It has been utilized as a part of conjunction with different methods to explain classification and prediction tasks. Various examinations have contrasted on naïve bayes and other machine learning, data mining procedures [13]. Naïve Bayes acutely completed with different procedures, for example, decision tree and neural networks. This may presumably be because of the way that specialists have not focused on the capacities of naïve bayes. In any case, this does not make naïve bayes less helpful meanwhile it beaded as rule different procedures when they were compared.

1.5.5 K-nearest neighbor (KNN)

K-NN is a technique for characterizing objects in view of nearest training data in the component space. It is a sort of instance-based learning. The k-closest neighbor algorithm is among the least complex of all machine-learning algorithms. However, the accuracy of the K-NN algorithm can be extremely debased by the nearness of noisy or irrelevant features, or if the component scales are not steady with their significance [18]. The classifier of closest neighbor depends on learning by relationship. The samples of training are depicted by attributes of n dimensional numeric. A k-nearest neighbor classifier searches the pattern space for the k training samples that are nearest to the anonymous sample at the point when unknown sample is given, [6] "Closeness" is defined as Euclidean distance, where the Euclidean distance between two points are: X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

The anonymous model is allotted the supreme common class between its k nearest neighbors. When k=1 the anonymous model is given to training sample class that is nearby its pattern space.

CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION:

This chapter will survey diverse types of research papers that are published in field of Data mining. Therefore, various classification techniques and technology that helps to find the problem in this field is explained below:

2.2 REVIEW OF LITERATURE

Fabricio Voznika et al. [6] explained that the approaches of data mining to reveal hidden patterns from huge amount of data. These hidden patterns can be utilized to forecast behavior of future. The accessibility of new algorithms of data mining, in any case, should be met with caution. Initially, these approaches are as good as with the collected data. Good data is the main prerequisite for good analysis of data. The subsequent step is to choose the best procedure to mine the data expecting great data is accessible. In any case, there are tradeoffs to consider while picking the appropriate data mining procedure to be utilized as a part of a specific application. There are different types of issues that are conductive to every technique. The "best" model is frequently found by experimentation: attempting distinctive technologies and algorithms. The data analyst should analyze or even consolidate accessible systems with a specific technique and accomplish the possible outcomes.

G.Kesavaraj et al. [4] applied the fundamental classification methods on various dataset. The main aim of this review was that it was give a survey of various methods of classification in area of mining. A few noteworthy types of classification method including Bayesian systems, KNN classifier and decision tree induction were explained in this paper. The main objective of these algorithms was produce more definite, clear-cut and exact simulation results. These are normally strong in modeling interactions. It was still very hard to prescribe any one method as better than others as per the given dataset. However, in this paper, J48 had the maximum accuracy i.e. 92.7624% having 333 instances incorrectly classified and 4268 instances had correctly classified. Finally, it was analyzed that for all sort of dataset single classification algorithm is not best.

Sagar S. Nikam [5] described different classification methods i.e. machine learning based and statistical based utilized as a part of data mining. It can be used as a portion of a wide range that directions techniques from various fields containing machine learning, Network interruption detection, artificial intelligence, measurements and pattern recognition for exploration of expansive volumes of information. The techniques of classification are ordinarily solid in presenting substitutions. Each of these techniques can be utilized as a part of different circumstances, as required where one has a tendency to be valuable while the other may not and the other way around. These algorithms can be executed on various sorts of informational indexes like offer market dataset, dataset of patients, financial dataset. Consequently, these methods demonstrated how an information could be resolved and gathered when another arrangement of new data is accessible. Every procedure has its own features and the limitations of these are like naïve bayes is easy to execute. Where SVM is highly accurate, classes need not to be linearly classified in KNN. According to Conditions, performance and required features can be chosen.

Yi Peng et al. [7] explained about classification techniques in Financial Risk Detection. Early detection of financial risks can support grantors of credit as well as organizations to build up proper approaches for credit items, decrease misfortunes and increment income. The classification techniques performance varied with various datasets. No single technique had been observed to be better over others for all datasets. The main idea of this paper was to give relative exploration of the capability of a choice of prevalent techniques of classification to forecast risk. The result of this can benefit institutes of finance to select suitable classifiers for their particular tasks. It was fail to perceive risk may cause grantors of credit and organizations serious losses of finance. This paper examined the execution of eight classification strategies i.e. Bayesian network, Naive bayes, SVM, KNN, C4.5, RBF, Linear logistics and RIPPER rule induction for risk analysis utilizing four genuine risk of finance related datasets. From which, Naïve Bayes and SVM are best for credit card application datasets and Bayesian Network accomplished great outcomes for bankruptcy datasets.

Mennat Allah Hassan et al. [8] presented that data mining can be utilized as a part of improving the nature of the services of medical obtainable through examining data as well as finding concealed patterns and associations that upgrade and even alter the behavior techniques received. In this paper, ten classification algorithms were applied on a patient's dataset obtained from a public hospital's database that contains patients both medical and personal information needed for diagnoses and treatment decisions. These algorithms analyzed using a data-mining tool and a comparative study was done to find the classifier that performed the best analysis on the dataset obtained using a set of eight performance metrics to compare the results of each classifier. According to the results, Bayes Net was the best classifier for this dataset with a True Positive Rate of 0.987, False Positive Rate of 0.002, Precision Rate of 0.988, Recall rate of 0.987, F-Measure of 0.988; ROC (Receiver Operating Characteristics Curve) area of 0.994 and the time taken for these results to be made was 0.0l. Thus, different medical institutes to compare different classification algorithms that can be applied on a medical data set for mining this data to discover new hidden patterns that can aid in the future treatment decisions can use these results. For the future work the data set used after it has been verified using Weka and its best classifier, it shall be used for adopting different Ontology aspects which is another technique of data mining to create different profiles for a set of diseases that can be shared amongst different medical personnel to aid in solving the interoperability issues.

Hong Yu et al. [9] explained about Individual credit risk evaluation is a vital and challenging data mining issue in financial analysis domain. This paper compared the adequacy of four data mining algorithms - logistic regression (LR), decision tree (C4.5), support vector machine (SVM) and neural networks (NN) by applying them to two credit datasets provided by UCI KDD Archive. The evaluation was done by examining the performance in light of classification accuracy, precision, and Type I error rate. Computer simulation shows that the classification effectiveness of the LR and SVM algorithms are best in general. In the condition of training with a small sample size, SVM algorithm is recommended because of its' higher robustness and generalization ability. Overall, the classification results of the C4.5 algorithm are acceptable. Given the better interpretational capability of the C4.5 algorithm, it is more appealing to business experts, and can be used to aid decision-making for banks in evaluating their future clients 'credibility. Further research is called for on how to combine the SVM algorithm with the C4.5 algorithm, integrating their robustness, generalization ability and the better interpretational capability, and apply them in individual credit risk evaluation.

Hamidah Jantan et al. [10] analyzed that how some talent management issues could be illuminated using classification and prediction methods in data mining. There are numerous data mining methods applied to the distinctive problem areas in human resource field of research into expand the skyline of academic and practice work on data mining in human resource. Other Data mining techniques such as Support Vector Machine (SVM), Fuzzy logic and Artificial Immune System (AIS) should also be considered for future work on classification techniques using the same dataset. In some cases, the attribute relevancy also reacts as a factor on the accuracy of the classifier algorithm. In the next experiments, the attribute reduction process should be implemented using other reduction techniques to confirm these findings. C4.5 classifier has the highest accuracy in the experiment; the accuracy of other decision tree classifier should also be tested in order to validate these findings. In addition, C4.5 classifier algorithm is the potential classifier in this experiment. Thus, this technique should be applied in the next prediction phase to construct classification rules. These generated classification rules can be used to forecast the potential academic talent. Lastly, the ability to uninterruptedly change and obtain new understanding of the classification and prediction in HR researches has thus, become the major contribution to HR data mining.

Zhenni Feng et al. [11] explained about trajectory data mining is beneficial to individual citizens. One can understand his or her movement behavior better through analyzing historical trajectories. Besides, trajectory data mining provides plenty of convenience to the public, e.g., route recommendation, real-time traffic information publication by transport agencies. However, people suffer from privacy breaches if their trajectories are collected and utilized inappropriately. Moreover, people are usually disturbed by commercial advertisements, which are possibly pushed in the name of personalized services. Trajectory data mining helps to reduce cost of supervision and management for the government and some organizations. In urban areas, trajectory data mining from vehicle trajectories provides an efficient and scalable method to monitor traffic condition of the whole city. Another example is to record illegal or irregular behavior, which is probably valuable to ascertain responsibilities later. For example, over speed can be inferred from trajectories. This evidence is valuable especially in roads without roadside cameras. Similarly, commercial organizations expect to cut down their costs in virtue of trajectory data mining. For example, RFID data, as a special kind of trajectories, indeed help to manage commodity stocks. Location acquisition technologies generate huge amount of trajectory data. Trajectory data, which track traces of moving objects, is typically represented by a sequence of timestamped geographical locations. A large amount of applications is created upon mining trajectory data. The survey reviews an extensive collection of existing studies in the proposed framework of trajectory data mining.

Navjot Kaur Walia et al. [12] explained that carbon plays a vital part in the environment for climate change. The absence and presence of carbon directly influences every single living being. Trees breathe in carbon and breathe out oxygen. CO2 is kept in numerous five-carbon pools of forests. The main objective of this paper was to build a system utilizing naive bayes approach that trains a model to classify forest on the basis of carbon stock and predict the level of carbon stock in the forest. The model is validated using dataset of the previous year data. Naive Bayes is a classifier that is used to build a model given some inputs and a label upon which the classification is required. In this case, first the model was created using the dataset. Then was tested using a testing dataset. Thus, the model classifies the data correctly by given the predicted values for label. The tool used for this experiment is Rapid Miner Tool that is a popular tool for analysis of data through data mining techniques. The model gives 81.50% accuracy and the testing dataset classifies two labels correctly out of the three given inputs.

Anshul Goyal et al. [13] explained about performance of naive bayes and J48 classification algorithm. Naive Bayes algorithm depends on possibility and J48 algorithm depends on decision tree. This paper set out relative evaluation of classifiers concerning financial institute dataset to boost maximum true positive rate and minimum false positive rate of defaulters instead of accomplishing just higher classification accuracy utilizing Weka tool. The outcomes on this dataset additionally demonstrate that the efficiency and accuracy of J48 and naive bayes was good. Bank data set had been taken from UCI repository having 300 instances and 9 attributes. J48 is a simple classifier technique to make a decision tree, efficient result had taken from bank dataset using weka tool. The results were about classification accuracy and cost analysis. J48 algorithm gives more accuracy for class gender in bank dataset having two values Male and female i.e. 52.67% and 52%.

Tina R. Patil et al. [14] explained about the evaluation of performance based on the correct and incorrect instances of classification using naïve bayes and J48 classification algorithm. Classification was used to classify the item according to the features of the item with respect to the predefined set of classes. The experiments results shown in this paper were about classification accuracy, sensitivity and specificity. The results also show that the efficiency and accuracy of J48 is better than that of naïve bayes on bank dataset. The author concluded that correct instances generated by J48 are 203 and Naïve Bayes are 184. This demonstrates that J48 was a simple classifier method to make a decision tree. Productive outcome had taken from bank dataset utilizing weka tool. Naive Bayes classifier likewise indicating great outcomes. Nonetheless, J48 gave more classification precision for class contract in bank dataset having two esteems Yes and No.

Md. Sajidur Rahman et al. [15] presented a comparative evaluation of classifiers Naive Bayes, Multilayer Perception, J48 Decision Tree and KNN in the context of household datasets to true positive rate and false positive rate of defaulters rather than achieving only higher classification accuracy using weka tool. This paper investigated and analyzed the existing raw data improved cook stoves (ICS) from different household information in Bangladesh; and also expedites Association Rule to extract some important information that can be helpful for future deployment, analysis and planning for sustaining efficient improve cook stove (ICS) program. The implementation of different classification algorithms and Apriori algorithm can be efficiently done to distribute the New ICS cook stove. Here multilayer perception model showed the highest accuracy 92.093% and specificity 96.975% and the J48 has the highest sensitivity using Weka environment. The dataset shows that, most of the household are stay under poverty line and majority of the people in the rural area suffers suffocation and the kitchen has carbon mark. Therefore, by using new ICS they have to contribute money for saving and improve their kitchen’s health condition

Huimin Xiang et al. [16] presented that the research trend of soil carbon stock has steadily increased from 2000 until 2014, with its research output significantly increasing from 2011 until 2013. These trends imply that soil carbon stock as an important research area for addressing climate change received greater attention. As the global climate change problem still exists and soil carbon stock research will continue to grow in the future. Except for the PR China, Brazil and India, research output was concentrated in the developed countries, such as USA, Germany, the UK, Canada and Australia; however, 133 countries throughout the world contributed to soil carbon stocks research, showing that climate change is a global problem with soil carbon stocks research continuing to gain worldwide popularity. Research had two main trends between 2000 and 2014. First, it focused on the impact of outside factors on soil carbon stocks, such as climate and vegetation, as well as human activities such as agriculture, forest management and so on. Second, concerning the dynamics and cycle of the soil carbon pool, research focused on both outside and inside factors. Research regarding soil carbon stocks clearly advanced human understanding of the relationships between climate change and soil carbon pool, though at times researchers did not reach a consensus on certain themes.

G.G.Gokilam et al. [17] explained about the performance of some data mining classifier algorithms named J48, Random Forest, Random Tree, REP and Naïve Bayesian classifier are evaluated based on 10 fold cross validation test. Diabetes is the most rapidly growing chronic disease of our time. In this paper, the author had taken diabetes and heart datasets related with their matching fields then apply the classification algorithm in diabetes heart dataset in weka tool finding weather people affected by diabetes are getting chance to get heart disease or not, output are evaluated as Tested Negative (No Diabetes), Tested Normal (Not affected), Tested High(affected).This research work proposed a new approach for efficiently predicting the diabetes heart disease from some medical records of patients. Dataset has designed with matching attributes applied in classification algorithms like J48, Random Tree, Random Forest, REP, and Naïve Bayesian in WEKA Tool. On this experiment classification wise J48 Produces highest accuracy, 95% apart from decision tree Naive Bayesian take minimum time (0.00 Seconds) to classify.

Surabhi Chouhan et al. [18] explained that improved Feature Selection and Classification technique is implanted on Benchmark Datasets such as Mushroom and Soyabean. The Proposed Methodology implemented is based on the Hybrid Combinatorial method of Applying PSO-SVM (particle swarm optimization) for the selection of Features from the Dataset and Then Classification is done using Fuzzy Based Decision Tree. Experimental results when performed on Various Datasets prove that the proposed methodology extracts more features as well as provides more accuracy as compared to existing methodologies. The result analysis shows the performance of the proposed methodology. The proposed methodology implemented here provides more accuracy for the classification of Datasets such as Soyabean or Mushroom Dataset. Various Experimental Results when performed on these dataset provides more generated rules and high selection of features using PSO-SVM algorithm and Fuzzy Decision Tree. Hence provides high Accuracy as compared to the existing methodology and less Error Rate and High Positive Rate. Although the methodology applied here provides efficient results as compared to the other existing techniques, but further enhancements can be done related to the execution time of the methodology as well as reducing the rules generated.

V. Rajeswari et al. [19] explained about different techniques of data mining classification methods were used to predict type of soil for example, Naive Bayes, JRip, J48. These techniques were useful to extricate the statistics from dataset of soil. Two Kinds of soil viewed in this as Red and Black. The two techniques were summarized in this paper i.e. Data Mining and farming Data Mining. The Kappa Statistics extended in the forecast. The JRip model produced highly reliable consequences of this information. Data Mining was used to improve the precision of classification of gigantic soil data collections for understanding the issues in Big Data, proficient strategies can be made. The comparison study of these three techniques like Naïve Bayes, JRip and J48 was predictable. The classification algorithm JRip provided improved result of this soil dataset with 98.18% accuracy. It was accurately characterized into most extreme number of examples contrasting with the other two. It was recommended that to predict soil types, JRip performed best results.

Simge Deniz et al. [20] explained about an empirical analysis of fuel consumptions and emission levels of passenger cars in context of application potential of big data mining. In this way, the dataset, which has been resulted from combining Euro 6 and technical specification data of Mercedes Benz, BMW and Audi automobiles, passenger cars have been used. Results of descriptive statistics and information about data have collected by different data mining tools. Usages have been compared and parameters have been derived by clustering algorithms aiming better result by categorizing of variables for upcoming analysis. The results obtained from various classification techniques i.e. Bayesian network, neural network and C5.0. From which C5.0 algorithm has given a better prediction for most of the cases like CO2 emission 90.2, CO emission 70.9 and emission of other particulates 96.1. Additionally, the importance of each parameter has been evaluated to predict its contribution on fuel consumption, on emission level and noise level. C5.0 algorithm has found the weight factor more effective on fuel consumption, fuel type on emissions and transmission on noise level. Engine capacity and weight have been the most important parameters for Bayesian network. Engine capacity also effects noise level more than other parameters. Results of neural network have shown that weight, transmission and fuel type of passenger cars have been the most important indicators for fuel consumption and for emissions. Transmission, wheel size and weight, respectively, effect noise level of vehicles at most. In addition, the results of this work will help improving of vehicle design phase in term of vehicle's environmental certification. On the other hand, embedding the results of this paper into mathematical models will improve vehicle routing problems and simulation-based evaluations.

Fahad Sheikh et al. [21] described about forecasting the weather using number of algorithms and different methods of mining, which was necessary to notify individuals and in advance prepare them about present and forthcoming condition of weather. A classifier was obtained that should have been utilized to forecast weather and also figure out the climate condition of particular area on the basis of accessible prerecorded information which supports assets and plan for the upcoming progressions. The analysis of Naïve Bayes, C4.5 and Decision Tree was done concurrently with dataset containing information about climate that was gathered over a time of 2 years for present utilization of data mining in climate forecast domain. The result discovered that the performance of C4.5 (J48) decision tree algorithm was much better than that of Naïve Bayes. The accuracy value of C4.5 was 88.2% with correct instances to classify them. Then again, Naïve Bayes demonstrated a poor result of 54.8% while classifying the instances. In this way, one might say that the performance of C4.5 method was superior to anything that of Naïve Bayes if there should be an occurrence of dataset managing climate.

HU Xiaoliang et al. [22] defined that different forms of buses in china have emissions of carbon dioxide at certain level. The data of Shanghai and Changzhou, which are two cities of china, were used in this paper. To estimate the bus emission of CO2 finds a way to verify it using the Vehicle Specific Power (VSP). It additionally provided a quantitative exploration of emission of CO2 per capita of various types of transports. This examination had discovered the indications to strengthen that "Transport need" system from the information and the figuring of CO2 outflow. With a significant occupancy rate in the general transport, transport paths and BRT transport fast travel frameworks improve execution in decrease of CO2 for the change of speed and higher volume for travelers. Among these three, the general transport had 62.35g emission of CO2 for every capita per km, while the information of transport path is 45.95. BRT had the best performance of CO2 outflow per capita of 31.14 for its high limit and selective street path. The productivity of CO2 reduction is reliable and significant for the courses with relentless and extensive traveler's stream.

Chairul Saleh et al. [23] explained about the model of SVM that was proposed for the prediction of consumption of carbon (CO2) discharge. The utilization of energy, for example, electrical vitality and consuming coal is input variable that influence specifically expanding the outflows of CO2 were directed to construct the model. The principle objective is to screen the emission of CO2 in light of the electrical vitality and consuming coal utilization from the procedure of production. Dataset was partitioned by techniques of cross-validation into 90 of training data and 10 of testing data. To locate the ideal parameters of SVM was utilized the experimentation approach on the trial by changing C parameters and Epsilon. The smallest error of this represents more accurately prediction. Experimentation approach was connected all together get a superior forecast demonstration with a lower error rates. The outcomes acquired demonstrate that the lower mistake (RMSE) value was 0.004 with ideal parameters for the SVM model of 0.1 for the C parameter and 0 for Epsilon. Forecast with high exactness can give data worried about CO2 discharges. It can be presumed that when the high precision of the forecast model, at that point the lower RMSE esteem must be obtained. By observing utilization of energy, it can enable the administrator to create policies or taking a choice in order to decrease the negative effect on the earth during the production procedure. In addition, the parameter of SVM model can be consequently preferred by incorporating improvement method, for example, genetic algorithm or optimization of particle swarm.

A. S. Galathiya et al. [24] explained about the comparison of best-known supervised classification techniques and it produced a basic correlation between supervised techniques like Decision Tree with Naïve Bayes, KNN, SVM, Neural Networks and Bayesian Classifier. The idea was not clear which classification-learning algorithm was better than other methods, however under which situations a specific strategy can altogether outflank others on an application issue that was given. This exploration work proposed C5.0 classifier that performs selection of feature, cross validation, diminished error pruning and model many-sided quality for unique C5.0 to lessen the optimization of proportion of error. The essential assignment of classification is to classify new and inconspicuous samples effectively. C5.0 is a classifier that gives effective classification in less time contrast with other classifier. Memory utilization is less in producing decision tree. The primary aim of research was identified with enhanced precision and create small decision tree. The precision was increased by 1-3 by the new framework. Therefore as the further scope, the execution was accomplished for the new elements like Feature Selection, Reduced Error Pruning, Cross Validation and Model Complexity. By actualizing the algorithm diversities utilizing RGUI with weka packages, the accuracy of classification was improved.

A. Sai Sabitha et al. [25] described that determination of an appropriate LO (learning objects) having an instructive worth, should be adapted towards convenience, which have importance to the learner. To upgrade the academic benefit of learning content and to fulfill the adapting needs of various student, joining of learning objects and OER (open educational resources) was done utilizing classification procedures of mining. The OERs were broadened learning content and are considered as Knowledge Objects in this examination. This was accomplished by conveying both LOs and improved learning content (OERs). Data mining applications has been broadly utilized as a part of an e-learning condition with mobile learning has gained its importance by providing a platform to learn, anywhere, anytime. Four-classification system was utilized to understand and recognized the appropriate OERs with LOs. The systems utilized are Decision tree, KNN, Neural networks and Bayes classifiers. The Naive Bayes classifier system depends on Bayesian theorem and is especially suited when the dimensionality of the sources is high and simple to build. It can regularly outflank more modern classification strategies like decision tree, which have the issues of over fitting, KNN that relies on coordinating of test objects with training objects, and neural network relies on upon the introduction parameters for training the network. The quick pattern of e-distributing and online networking appears there can be a combination of open and exclusive learning objects. This work appeared the basic issue in the recognizable proof of learning assets that a specific group of learners requires keeping in mind the complete their assignments and the careful coordination of OER can go about as a key asset, in the learning procedure. Classification depended on instruction level and should likewise be possible on different qualities as subject, age of the learner and so on. This might be reached out by incorporating different sources of information removed from KMS and further can be grouped using ensemble methods.

Vincenzo Manzoni et al. [26] summarized a technique to evaluate progressively the CO2 discharges utilizing inertial data assembled from cell phone sensors. A calculation distinguishes transportation modes utilizing decision trees as a supervised technique. This application can be keep running on standard cell phones for drawn out expanses of time and can work straightforwardly. Firstly, make utilization of a current platform cell phones that is broadly received, this technique had the capability of extraordinary information gathering of mobility designs. This technique utilizes information from the gadget's accelerometer, while GPS information and online guide queries are utilized just sparsely. Moreover, it enable the mobile to be ordinarily carried in a client's pocket, without the requirement for particular fitting as far as position and introduction. At last, the information is made significant to the client by changing over it into emission of CO2 as an element of method of transportation, remove, and consumed calories for health monitoring. The technique’s components are the fluctuation and the FFT (Fast Fourier Transform) coefficients of the aggregate quickening measured by a cell phone's accelerometer. The outcome demonstrate that the accuracy of the algorithm is 82.14%. The strategy does not depend on a particular introduction of the telephone or on information from GPS as past research, works do. Nevertheless, the GPS and even online maps administrations can enhance the grouping exactness assist on. Shockingly, GPS readings and online inquiries are vitality-escalated forms. Regardless, the GPS gadget must be misused to register the voyaged distance, an obligatory contribution for processing the CO2 outflows.

Mohamed Beidari et al. [27] described about the analysis of input-output to find the 18 accumulated areas for the years 1995, 2000, 2005, 2010 and 2012 interconnection. The received analysis of multiplier to evaluate the aggregate ecological effects identified with the business linkages for the year 2012 in South Africa, with an emphasis on the connection between the power division and whatever remains of the economy. In the first place, the effects of linkage were figured using the PIOD, which means reverse linkage, and SIOD means forward linkage. The results of power segment has a weak linkage both in reverse and forward linkages are under 1 with others segments, which implies it is generally autonomous of different divisions. In another words, it does not prompt and empower financial development by IOA. Two parts, for example, Chemical, Petrochemical Industries, and Basic Metals were establish as vital segments in economy in 1995, 2000 and 2012 years. Chemical and Petrochemical Industries was the most vital segment in SA. The multiplier connected to outline the 18 areas vitality and CO2 discharges forces for 2012 year. The outcomes plainly demonstrated that Water Supply, Electricity, Gas. Other Industries like Transport; Basic Metals and Residential were the main five vitality areas. The power segment was the principle coordinate money related vitality shopper and carbon dioxide producer, and subsequently it is the most predominant source as far as vitality. Besides, the consequences of multipliers analysis demonstrated that most aggregate of indirect vitality utilization and CO2 outflows were higher than energy utilization. This implies both indirect energy utilization and CO2 outflows that affects South Africa energy utilization. In this manner, the South Africa government should plan to execute common sense systems to diminish the energy utilization intensity and additionally the CO2 emanations power. In light of the consequences of this paper, an assortment of recommendations which can be valuable to enhance the power segment's linkage impacts to wind up noticeably a key area for South Africa, and decrease its immediate utilization of energy and discharge of carbon dioxide, have been prescribed in this segment. At last, this examination demonstrated that information yield analysis could be an extremely helpful apparatus for governments, as it gives significant data about the working and structure of the linkages among areas in the national level economy, and access to both immediate and aberrant impacts identified with vitality utilization and discharges of co2. The multipliers examination connected in this exploration gives a decent understanding of the interconnectedness among businesses for a government to assess the profiles of the immediate and indirect impacts among ventures with respect to their energy utilization and emission of CO2. Hence, it can be a beneficial device to help policymakers in explaining suitable financial approach and enhancing vitality arrangement makings.

Reina Kawase et al. [28] presented that long haul situations for atmosphere adjustment in European nations, which go for a 60–80 decrease in CO2 emission by 2050, were examined by deteriorating the adjustment in CO2 outflow, and the rates were contrasted and the pace of recorded change. At the point when GDP development was thought to be the same with respect to the previous 10 years, the situations for EU nations have the same principle countermeasures: Aggregated intensity of power and carbon must be enhanced at a pace of more than 2–3 times their verifiable change. These rates required to be conserved for a long time from here on. The decline of intensity carbon is accomplished by constant non-renewable energy source use with CCS, or a large portion of the rate is accomplished by petroleum product utilize decrease and the rest of atomic power and/or renewables. Contrasted with Japan's authentic change rates, the required change rates of accumulated power intensity and carbon power ought to end up noticeably 3–5 overlap. For the Japan's long haul atmosphere adjustment situation, the required change rate for a lessening in carbon power and a change in amassed intensity was broke down with the presumption of GDP development, the measure of CCS, and vitality transformation proficiency. The essential issue is the way Japan joins the adjustments in these lists: depending on effectiveness changes in vitality administrations and gadgets, or the troublesome decisions, for example, atomic power and CCS.

Jyoti Soni et al. [29] explained about the issue of summarizing and constraining distinctive data mining algorithms utilized as a part of the medical field. The essential objective of this review was used distinctive algorithms as well as combination of different target qualities for effective prediction of heart attack utilizing mining procedures. Through 15 properties recorded in addition with basic data mining strategy different techniques e.g. ANN, Time Series, Clustering and Association Rules, soft computing methodologies and so on can be consolidated for prediction of heart attack. On the same dataset, the consequence of prescient data mining method reveals that Decision Tree outflanks and at some point Bayesian classification was having comparative exactness as of decision tree in any case, other prescient systems like KNN, Neural Networks, classification in light of grouping are not performing great. The second conclusion is that the precision of the Decision Tree and Bayesian Classification also improves in the wake of applying genetic algorithm to diminish the genuine data size to get the perfect subset of attribute satisfactory for prediction of heart disease. The proposed work can be additionally improved and expanded for the computerization of prediction of Heart ailment. Genuine data from Health mind associations and offices should be gathered and all the available systems will be compared about for the optimum accuracy.

Milan Kumari et al. [30] analyzed that various techniques of data mining that can be utilized for the distinguishing proof and avoidance of cardiovascular disease amongst patients. Four classification strategies were contrasted for prediction of cardiovascular sickness i.e. rule based RIPPER procedures, decision tree, Artificial neural networks and Support Vector Machine in this paper. These procedures were associated on premise of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. According to outcomes, rates of error for RIPPER was 02.756, Decision Tree was 0.2755, ANN was 0.2248 and SVM was 0.1588 individually. The exactness of RIPPER, Decision Tree, ANN and SVM were 81.08, 79.05, 80.06 and 84.12 separately. This study demonstrated that SVM predicts cardiovascular disease with least error rate and most extreme precision from other four strategies. This study demonstrated that SVM show ended up being best classifier for cardiovascular illness expectation. In future, this expect to enhance execution of these fundamental techniques of classification by making meta model, which will be utilized for prediction of cardiovascular disease in patients.

CHAPTER 3
SCOPE OF STUDY

The scope of the study will focus on the correct analysis of worldwide emission of carbon dioxide. The scope of this research is as follows: To implement classification approaches as automatic or real time classification tools which may useful for experts to identify the increase of carbon dioxide so that it may help to predict environment changes. In these days, the ecological study of carbon is a key concern. Carbon cycle plays important role for life cycle on earth. It is acknowledged as one of the supreme dominating greenhouse gas (GHG). Each country emits different amount of carbon dioxide each year from different sources including burning of oil, coal and gas as well as deforestation. A rank is also given to each country based on statistical measures of education, poverty, life expectancy and income levels know as human development index (HDI) each year. The main aim of this study is to predict the level of development for each country based on human development index and carbon emission each year. Also find the best classification technique from various classification algorithms of data mining that are used to analyze the results on year wise data set, it is concluded that tree based algorithms are more accurately able to classify the data.

CHAPTER 4

PRESENT WORK

4.1 PROBLEM FORMULATION

The research problem is a crucial part of any research activity. If nature of the problem is clear that it is very easy to solve the problem.

Carbon emission represents major temperature change worldwide. Each year the overall carbon emission is rising. By looking at previous work it is observed that lot of working is done in the field of electronics and chemical structures but less amount of work is done is computer science field. Another factor besides Co2 emission is human development index, which includes factors like education, poverty and lifespan to rank each country. The problem of the study to categorize countries based on HDI and Co2 emission using data mining algorithms.

The purpose of this work is to determine the performance of various algorithms of classification on basis of different parameters. Classification method is supervised and assigns objects into sets of predefined classes. There are diverse sorts of classification approaches being utilized in data mining such as rules trees and function. The primary objective of classification is to accurately calculate the value of each class variable. This classification method involves two stages i.e. training and testing. The first step is to build the model from the training set, i.e. casually samples are carefully chosen from the data set. In the second step, the data values are allotted to the model and validate the model's accuracy. In the base paper, only Naïve Bayes is used solely on Co2 forest data without any other attribute. In this paper, several algorithms on same dataset are applied to select the best one with high accuracy.

4.2 OBJECTIVES OF STUDY

Any task without sound objectives is like tree without roots. In the same way during any research undertaken, first objectives of research study are determined and then next steps are taken in order to proceed further. A research study may have many objectives but all these objectives revolve around one major objective, which is the focus of study.

The main objective of the study is to analyze the emission and impact of carbon dioxide worldwide. The following objectives have been set forth. They are:

i. To collect and preprocess data available from different sources and convert it into valid training set using excel VBA and Replace Missing Value Filter (RMVF).
ii. To implement various data mining algorithms on carbon dataset for evaluating performance and country categorization.
iii. To analyze variation in performance using Cross Validation and Split Validation Techniques.
iv. To compare and analyze the results from different techniques on basis of Accuracy, Precision and recall.

4.3 RESEARCH METHODOLOGY

The following are main steps to perform and find the best classification technique among number of techniques.

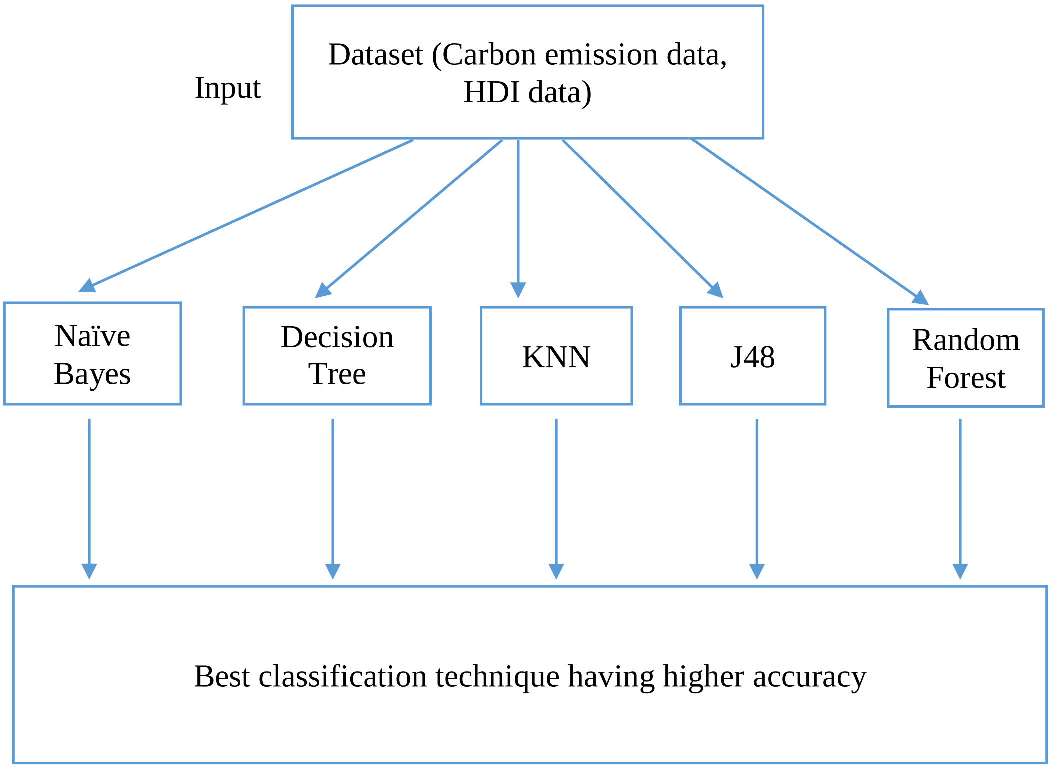
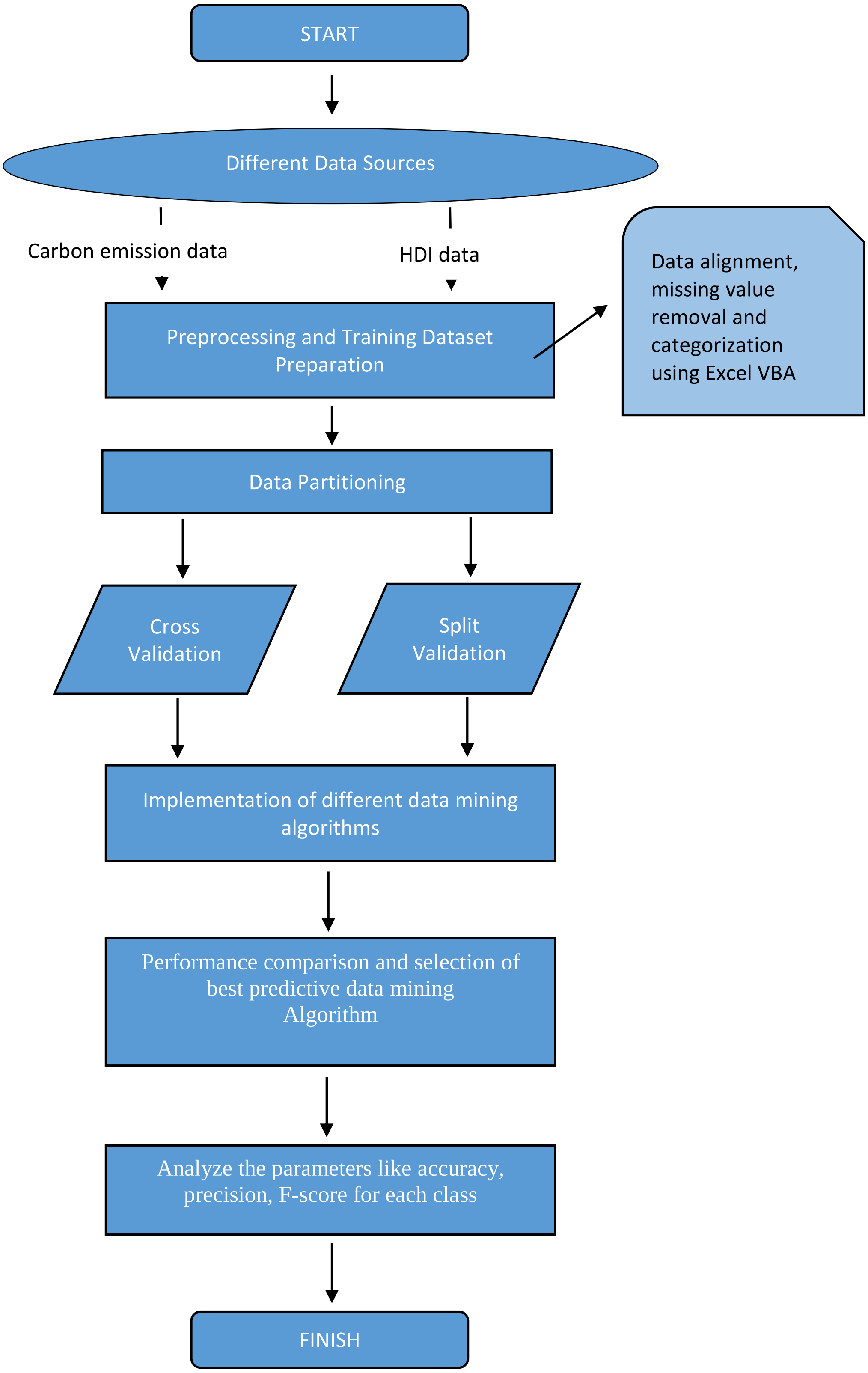

Input
Dataset (Carbon emission data, HDI data)
Naïve Bayes
Decision Tree
KNN
J48
Random Forest
Best classification technique having higher accuracy

Figure 4.1: Processing steps of proposed model

4.3.1 Proposed Method



START
Different Data Sources
Carbon emission data
HDI data
Preprocessing and Training Dataset Preparation
Data alignment, missing value removal and categorization using Excel VBA
Data Partitioning
Cross Validation
Split Validation
Implementation of different data mining algorithms
Performance comparison and selection of best predictive data mining Algorithm
Analyze the parameters like accuracy, precision, F-score for each class
FINISH

Figure 4.2: Working of proposed model

This section explains all the tools, techniques and steps carried out to perform classification on country wise carbon emission and human development index dataset for doing a comparative analysis between the algorithms based on accuracy. This approach helps identify the appropriate class for each country based on conditions applied. The work is done on Windows 10 operating system with i7 processor and 16gm of ram. Rapid Miner Studio is used to apply algorithms on dataset. Microsoft excel is used to store the data and excel VBA programming is applied on data to convert it into a valid training dataset. Every step is explained in detail below.

1. The collected data for carbon emissions and human development index for each country is available on web freely but not at single location so for that data from different sources were collected and arranged accordingly in excel sheet. The year wise data is available and for this research the latest dataset of year 2016 is collected. To replace any missing value from dataset to avoid results conflict, the Replace Missing Value filter is applied on dataset. To divide the data into three classes, four conditions were made and excel VBA is used to automatically calculate a class for each tuple. Conditions are as follows.

    - If HDI is above 50 and Carbon emission is below 5000 metric tons then country development is average.
    - If HDI is above 50 and Carbon emission is above 5000 metric tons then country development is negative.
    - If HDI is below 50 and Carbon emission is below 5000 metric tons then country development is positive.
    - If HDI is below 50 and Carbon emission is above 5000 metric tons then country development is average.

2. For data partitioning two techniques were applied on dataset which helps us analyze the variation in performance for each algorithm. Both the operators are available is Rapid Miner Namely Cross-Validation and Split-Validation.

    - **Cross-Validation** – This operator is used to automatically divides the whole dataset into training and testing dataset. It works as nested operator with two sub processes. In first phase the model is trained using testing dataset and then this model is applied on testing dataset for calculating the performance of a model. The inputs require are number of folds and type of sampling. For instance, if k folds are given the k-1 folds are used for training the model and remaining one for testing. Sampling can be done as linear, shuffled, stratified or automatic.

- **Split-Validation –** This operator works somewhat similar to Cross Validation bus with some difference in dividing the dataset. Instead of several iterations or the number of folds it splits up the data into only two sets i.e. training and testing and the proportion of each is controlled using split ration where 1 means take all the data as training and 0.9 means 90 for training and rest 10 for testing.

3. In Rapid Miner the two operators for splitting the data needs an algorithm to be used to train and test the model. For that reason, different data mining algorithms including Naïve Bayes, K-Neatest Neighbor, Decision Tree, Random Forest and W-J48 are used to make different models. After getting the results, the confusion matrix for each model is inserted into excel sheet to do further calculations. As our data is polynomial and not binomial the precision and F-scores cannot be calculated directly. The result of these factors is obtained on basis of individual class where correct predicted results are considered as one class and wrong predicted results makes second class.

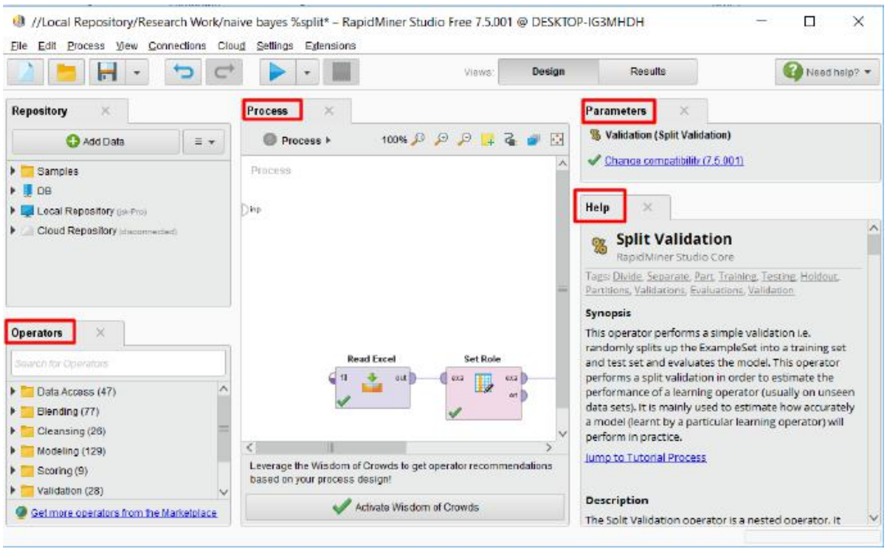
CHAPTER 5

RESULTS AND DISCUSSION

5.1 TOOLS DESCRIPTION

Rapid Miner

Rapid Miner is a GUI based data science tool developed by same the company with same name as tool. It has all the basic features for machine learning, data or text mining and data preparation tasks. It comes as a free version for individual use like students and as a commercial version used by business and commercial applications. Rapid Miner is may not include all the features like Weka but has a very user-friendly interface. The tool is designed to proceed the task using template-based framework without writing code hence producing results fast and less prone to errors. It has different sections to divide the related things in a group. All the functions used for data extraction, transformation and loading are called operators and it also gives brief introduction, the inputs and outputs necessary and example for each operator. Most of the operators have parameters, which can be changed using GUI interface without changing any code. The scripts from R and Python can also be used to include more algorithms within the tool. The diagram below shows the basic Rapid Miner interface.


//Local Repository/Research Work/naive bayes %split* – RapidMiner Studio Free 7.5.001 @ DESKTOP-IG3MHDH
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Need help?
Repository
Add Data
Samples
DB
Local Repository
Cloud Repository
Operators
Search for Operators
Data Access (47)
Blending (77)
Cleansing (26)
Modeling (129)
Scoring (9)
Validation (28)
Get more operators from the Marketplace
Process
Process ▸
100%
Read Excel
Set Role
Activate Wisdom of Crowds
Parameters
Validation (Split Validation)
Change compatibility (7.5.001)
Help

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Split Validation

RapidMiner Studio Core

Tags: Divide, Separate, Part, Training, Testing, Holdout, Partitions, Validations, Evaluations, Validation

Synopsis

This operator performs a simple validation i.e. randomly splits up the ExampleSet into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately a model (learnt by a particular learning operator) will perform in practice.

Jump to Tutorial Process

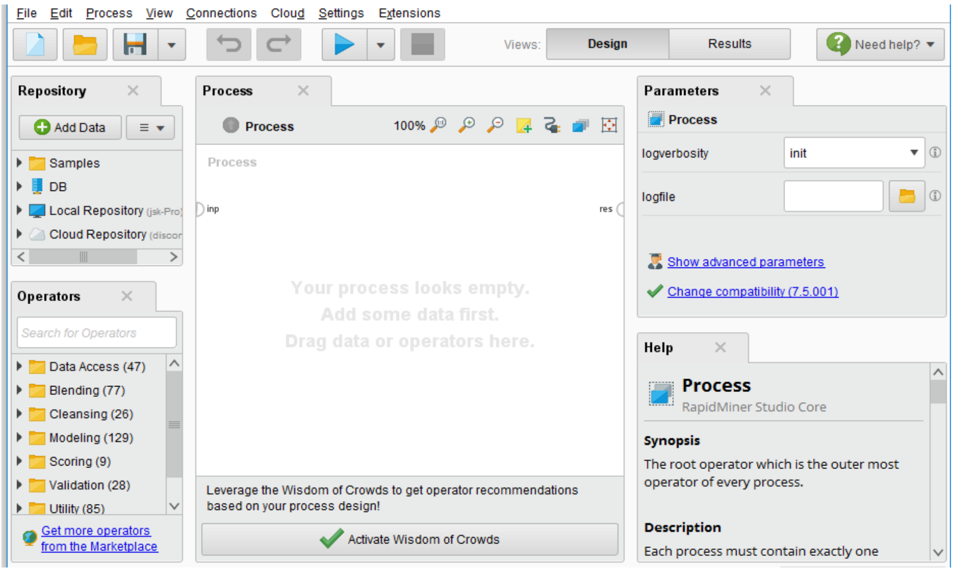
Description

The Split Validation operator is a nested operator. It

Figure 5.1: Interface of Rapid Minor tool

This section will explain the Rapid Miner workflow used for this research.

1. On opening the Rapid Miner, the process area remains empty as no operator is still added to it.



Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

logverbosity: init

logfile:

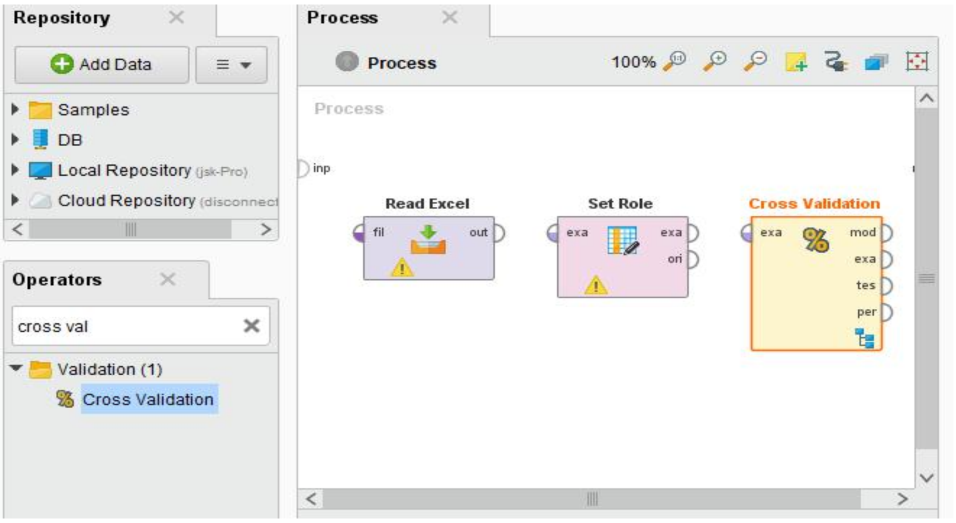
Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description
Each process must contain exactly one

Figure 5.2: Empty process

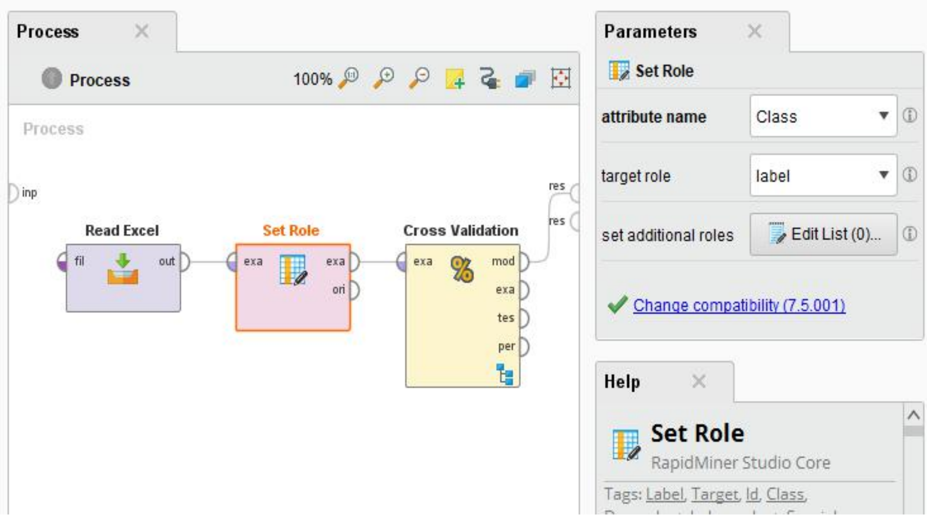
2. To perform several tasks like reading data from Excel sheet, setting the role for Class attribute and data partitioning some operators need to be dragged from operator area to process area. Each operator can be searched if the name is already known.



Search for operators: cross val

Figure 5.3: Adding operators to process area

3. The yellow exclamation marks on operator shows that no value is assigned to them. For that each operator can be selected one by one by left click and the parameters corresponding to selected operator will appear in parameter section. Dataset excel file is inserted in read excel operator and class attribute of dataset is used as label. Also to link the operators, click and drag from one port to another will make a link between operators.



Set Role

attribute name: Class

target role: label

set additional roles: Edit List (0)...

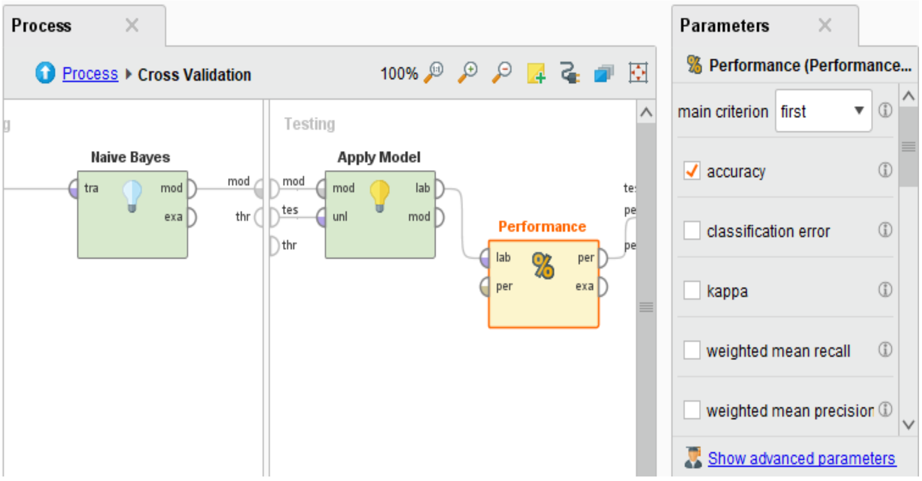
✓ Change compatibility (7.5.001)

Set Role
RapidMiner Studio Core

Tags: Label, Target, Id, Class,

Figure 5.4: Linking operators

4. To train and apply model, cross validation operator is further expanded by double clicking. It shows two areas for training and testing. For training section a single algorithm operator is selected. In testing section, the trained model is applied to calculate the performance.



Performance (Performance...

main criterion: first

☑ accuracy

□ classification error

□ kappa

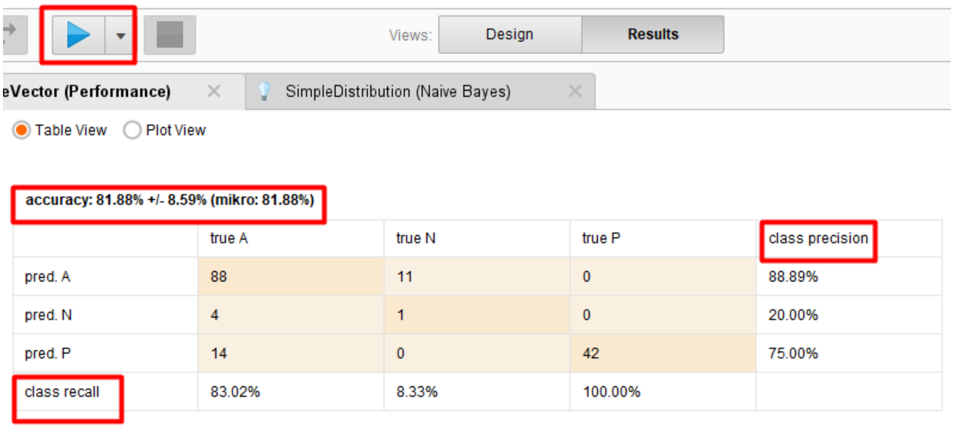
□ weighted mean recall

□ weighted mean precision

Show advanced parameters

Figure 5.5: Apply model to process

5. After all the steps if everting is connected correctly the by pressing the run process button at top toolbar will show the results of trained model.



accuracy: 81.88% +/- 8.59% (mikro: 81.88%)

	true A	true N	true P	class precision
pred. A	88	11	0	88.89%
pred. N	4	1	0	20.00%
pred. P	14	0	42	75.00%
class recall	83.02%	8.33%	100.00%	

Figure 5.6: Result window

5. 2 RESEARCH OUTCOMES

This section explains the results and conclusion of thesis work. Firstly, it explains all the performance evaluation parameters that are used for comparative analysis and after those experimental results are explained below.

5.2.1 Performance Evaluation Parameters

Performance of applied algorithmic models are measured on basis of several parameters like accuracy, precision and sensitivity (recall). To understand these parameters for this research some terms are defined below.
Three Classes are defined:

Average – When HDI is less than 50 and Co2 is greater than 5000 or HDI is greater than 50 and Co2 is less than 5000.

Negative – When HDI and Co2 both are greater than 50 and 5000 respectively.

Positive – When HDI and co2 both are less than 50 and 5000 respectively.

True positive (TP) = the number of countries identified correctly.
True negative (TN) = the countries not belong to current class are correctly rejected.

False positive (FP) = the countries belong to other classes assigned to current class.
False negative (FN) = the countries belong to current class are assigned to other classes.

1. Accuracy: The accuracy defines the closeness of predicted value to known value. In this case accuracy is to separate classes correctly. The mathematical formulae for accuracy is as follows.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

2. Precision: The precision is the fraction of relevant countries retrieved among all the retrieved countries. It contains countries identified incorrectly. Below formulae is used to calculate precision of given class.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

3. Recall: The recall is the fraction of relevant countries that are identified over total relevant countries in dataset, which are false negatives. Formulae is as follows.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

5.2.2 Experimental Results

For all the models applied are the data is divided into two sets using Rapid Miner Inbuilt Operators cross validation and split validation. Training and testing data are split in ration 80:20 which means 80 tuples from whole dataset will be used to train the model and rest to test the model. Training and testing data are remained same for all the algorithmic models. The performances of all models are evaluated in terms of accuracy, prediction and recall for each individual class.

Various classification algorithms are applied using Rapid Miner Studio. The table below show a comparison of accuracy and sensitivity using cross validation as well as split validation for individual class (average, negative, positive) using algorithms (Naïve, KNN, DT, J48, RF). After taking average of accuracy and sensitivity the DT algorithm is giving highest percentage of 99.51, followed by J48 also a form of decision tree gives percentage of 97.75 then RM gives about 10 less output i.e. 88.29. The KNN and Naïve gives worst results with percentage of 80.51 and 72.51 respectively.

Table 1: Comparative results of different classification techniques.

Classification Methods			**Naïve**	**KNN**	**DT**	**J48**	**RF**
Accuracy	Cross Validation (10 fold)	a=avg	81.88	80.00	98.75	98.75	96.25
		b=neg	89.73	95.52	100.00	100.00	98.09
		c=pos	90.34	83.12	98.75	98.75	97.41
	Percentage Split (80-20)%	a=avg	68.75	79.17	100.00	95.83	93.75
		b=neg	75.86	95.00	100.00	100.00	100.00
		c=pos	88.00	82.61	100.00	95.83	93.75
Sensitivity	Cross Validation (10 fold)	a=avg	88.89	83.64	99.06	98.15	96.30
		b=neg	20.00	75.00	100.00	100.00	100.00
		c=pos	75.00	71.05	97.62	100.00	0.00
	Percentage Split (80-20)%	a=avg	91.67	84.38	100.00	100.00	95.00
		b=neg	25.00	66.67	100.00	100.00	100.00
		c=pos	75.00	70.00	100.00	85.71	88.89
Average (%)			**72.51**	**80.51**	**99.51**	**97.75**	**88.29**

On comparing all the models solely on the basis of accuracy it is analyzed that decision tree algorithm gives 100 accuracy using split validation of ratio 80:20 and 98.75 using cross validation with 10 folds. The Naïve Bayes shows an accuracy percentage of 68.75 which is the lowest value among all the compared models.

Table 2: Accuracy of different classification techniques.

Classification Method		**Naïve**	**KNN**	**DT**	**J48**	**RF**
Accuracy	Cross Validation (10 fold)	81.88	80.00	98.75	98.75	96.25
	Percentage Split (80-20)	68.75	79.17	100.00	95.83	93.75

CHAPTER 6
CONCLUSION

Although there are many researches making advancements on environmental topics including carbon emission, which is responsible for temperature change but few researches, is done in field of computer science. The data on emissions is available freely over the web but without using any data mining, those are just bunch of numbers. The Human development index does not include any effect of carbon emissions by human activities, which becomes the focus of this research. It is analyzed that even countries with lower ranks producing high amount of carbon marks them as negatively developed countries. On comparing different models, decision tree gives higher accuracy using percentage split among all other classification based techniques. All tree based algorithms i.e. Decision tree, Random forest and J48 produced better results than rest of classification algorithms used.

REFERENCES

[1] S. K. Mahajan, “Human Development Index – Measurements , Changes and Evolution,” *NIRMA Univ. Int. Conf. Eng. NUiCONE-2013*, pp. 1–5, 2013.

[2] M. Ramageri, “Data Mining Techniques and Applications,” *Indian J. Comput. Sci. Eng.*, vol. 1, no. 4, pp. 301–305, 2010.

[3] M. Gera and S. Goel, “Data Mining -Techniques, Methods and Algorithms: A Review on Tools and their Validity,” *Int. J. Comput. Appl.*, vol. 113, no. 18, pp. 22–29, 2015.

[4] G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” *2013 Fourth Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–7, 2013.

[5] S. S. Nikam, “A Comparative Study of Classification Techniques in Data Mining Algorithms,” *Orient. J. Comput. Sci. Technol.*, vol. 8, no. 1, pp. 13–19, 2015.

[6] F. Voznika and L. Viana, “Data mining classification,” pp. 1–6, 1998.

[7] Y. Peng and G. Kou, “A Comparative Study of Classification Methods in Financial Risk Detection,” *Ncm 2008 4Th Int. Conf. Networked Comput. Adv. Inf. Manag. Vol 2, Proc.*, pp. 9–12, 2008.

[8] M. A. Hassan, M. Transport, E. A. Forces, E. M. R. Hamed, and M. Transport, “A Comparative Study of Classification Algorithms in E-Health Environment,” pp. 42–47, 2016.

[9] H. Yu, X. Huang, X. Hu, and H. Cai, “A comparative study on data mining algorithms for individual credit risk evaluation,” *Proc. - 2010 Int. Conf. Manag. e-Commerce e-Government, ICMeCG 2010*, pp. 35–38, 2010.

[10] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Potential data mining classification techniques for academic talent forecasting,” *ISDA 2009 - 9th Int. Conf. Intell. Syst. Des. Appl.*, pp. 1173–1178, 2009.

[11] V. Tanuja, “A Survey on Trajectory Data Mining,” vol. 4, no. 10, pp. 195–214, 2016.

[12] N. K. Walia, P. Kalra, and D. Mehrotra, “Prediction of carbon stock available in forest using naive bayes approach,” *Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016*, pp. 275–279, 2016.

[13] A. Goyal and R. Mehta, “Performance comparison of Na??ve Bayes and J48

classification algorithms," *Int. J. Appl. Eng. Res.*, vol. 7, no. 11 SUPPL., pp. 1389–1393, 2012.

[14] T. R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," *Int. J. Comput. Sci. Appl. ISSN 0974-1011*, vol. 6, no. 2, pp. 256–261, 2013.

[15] S. Rahman, "Carbon Emission Measurement In Improved Cook Stove Using Data Mining," pp. 83–86, 2017.

[16] H. Xiang, J. Zhang, and Q. Zhu, "A scientometric analysis of worldwide soil carbon stocks research from 2000 to 2014," *Curr. Sci.*, vol. 109, no. 3, pp. 513–519, 2015.

[17] G. G. Gokilam and K. Shanthi, "Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset," *Compusoft*, vol. 5, no. 3, pp. 2074–2079, 2016.

[18] S. Chouhan, "An Improved Feature Selection and Classification using Decision Tree for Crop Datasets," vol. 142, no. 13, p. 8887, 2016.

[19] V. Rajeswari and K. Arunesh, "Analysing soil data using data mining classification techniques," *Indian J. Sci. Technol.*, vol. 9, no. 19, 2016.

[20] S. Deniz, H. Gökçen, and G. Nakhaeizadeh, "Application of Data Mining Methods for Analyzing of the Fuel Consumption and Emission Levels," *Int. J. Eng. Sci. Technol.*, vol. 5, no. 10, pp. 377–389, 2016.

[21] F. Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun, "Analysis of data mining techniques for weather prediction," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016.

[22] H. Xiaoliang and C. Chuan, "Carbon dioxide emission comparison of different forms of bus in China," *2013 IEEE Elev. Int. Symp. Auton. Decentralized Syst.*, pp. 1–4, 2013.

[23] C. Saleh, N. R. Dzakiyullah, and J. B. Nugroho, "Carbon dioxide emission prediction using support vector machine," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 114, p. 012148, 2016.

[24] A. Galathiya, A. Ganatra, and C. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 2, pp. 3427–3431, 2012.

[25] A. S. Sabitha, D. Mehrotra, A. Bansal, and B. K. Sharma, "A naive bayes

approach for converging learning objects with open educational resources," *Educ. Inf. Technol.*, vol. 21, no. 6, pp. 1753–1767, 2016.

[26] V. Manzoni, D. Maniloff, K. Kloeckl, and C. Ratti, "Transportation mode identification and real-time CO2 emission estimation using smartphones: How CO2GO works," *Work*, pp. 1–12, 2011.

[27] M. Beidari, S.-J. Lin, and C. Lewis, "Multiplier Effects of Energy Consumption and CO2 Emissions by Input-Output Analysis in South Africa," *Aerosol Air Qual. Res.*, vol. 17, no. 6, pp. 1566–1578, 2017.

[28] R. Kawase, Y. Matsuoka, and J. Fujino, "Decomposition analysis of CO2 emission in long-term climate stabilization scenarios," *Energy Policy*, vol. 34, no. 15, pp. 2113–2122, 2006.

[29] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[30] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," *Ijcst*, vol. 4333, pp. 304–308.