Performance evaluation of Classification algorithms based on Weather Forecasting Dataset

Ajimol C,Mphil Scholar, Nanjil Catholic College of Arts and Science, Kaliakkavilai, Manonmaniam Sundaranar University, Tirunelveli 627 012.

Dr.R.Kavitha Jaba Malar, Assistant Professor,Department of Computer Science, Nanjil Catholic College of Arts and Science,Kaliakkavilai.

Abstract— **Data Mining is the process of extracting hidden knowledge from a large volume of data [7]. In this paper different types of classification algorithms including Naive Bayes, Logistic Regression, Random Forest and Support Vector Machine (SVM) algorithms have been comparatively tested to find the optimum algorithm for classification. Commonly a classification technique follows three approaches Machine Learning [9][10], Neural Network and Statistical for classification. This research analyzed the performance evaluation of popular classification algorithms used by machine learning systems. The main aim of the classification algorithm is to produce a good classification model which classifies the data accurately based on the training data set. Each classifier was tested with respect to accuracy, performance and execution time using the measured values. The experiments were carried out using the WEKA data mining tool, which includes the implementation of machine learning algorithms.**

Keywords*— *Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine (SVM), Weka Tool.

I. INTRODUCTION

Data mining is the process of discovering hidden information from large volume of data [7], which includes Anomaly detection, Association rule learning, Clustering, Classification and Regression which is used in various sectors such as Finance, Healthcare, Intelligence, Telecommunication, Sales and Marketing, E-commerce, Biological Data Analysis, Crime Agencies etc. Data classification plays a very important role in data mining. In this comparative study different characters of Naive Bayes, Logistic Regression, Random Forest and Support Vector Machine (SVM) algorithms were compared using Weather forecasting dataset from Indian Meteorological Department Website. The WEKA open source data mining software is used to evaluate the performance of the different supervised learning algorithms. WEKA measure parameters like Correctly Classified Instances (CCI), TP rate, FP rate, Precision, Recall, F-Measure, Root Mean Squared Error, Incorrectly Classified Instances (ICI) and build time. These outputs are gathered by using k-fold cross-validation method. Figure 1 explains the process flow diagram of the proposed method which includes data selection, pre-processing, transformation, data mining and evaluation.

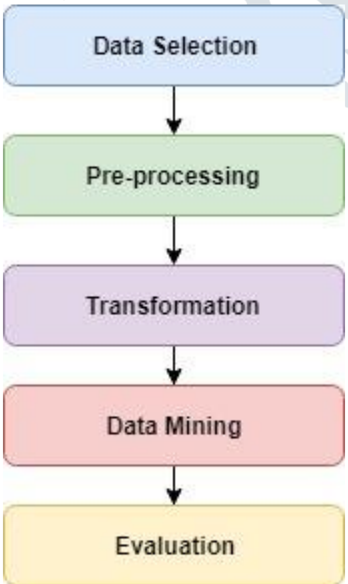

Fig 1: Process of Proposed Methodology

The rest of the paper is organized as follows. Section II covers the proposed methodology. Section III covers experimental results. Finally, in section IV, we conclude the comparative results.

II. PROPOSED METHODOLOGY

Open source Weka tool is used for this research Four different types of classifiers are used to perform comparative study Naive Bayes, Logistic Regression, Random Forest and Support Vector Machine (SVM. TP Rate, FP Rate, Precision (P), Recall (R), FMeasure (F) and accuracy of each classifier is calculated.

A. Data Pre-Processing

Data preprocessing is a vital step for supervised learning. Pre-processing refers to the changes applied to data sets before applying it to the model that convert the raw data into a clean data set. For achieving accurate results from the model in supervised learning algorithms the format of the data has to be in a proper format. The Weather Forecasting Dataset contains many imperfect data like missing certain fields to be filled by patients due to emergency cases, in some context collected data will be noisy. Filling the missing value in Weather Forecasting data is a very difficult task. Improperly handling the missing values will produce inaccurate results. In this comparative study multifilter is used for data preprocessing. Fig 2 shows the architecture flow of data preprocessing.

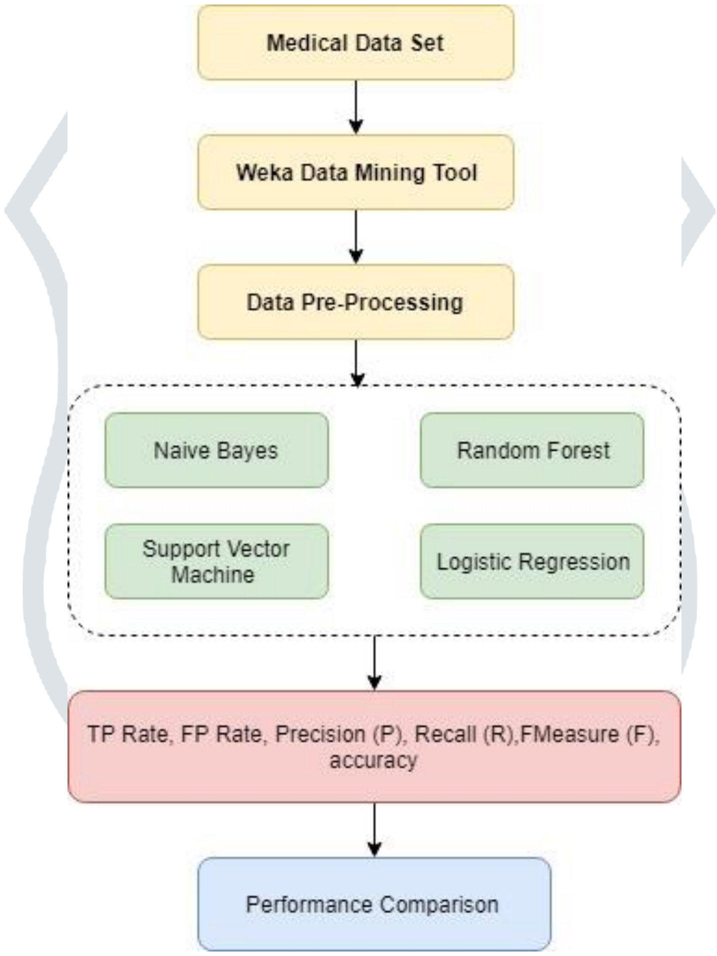

Fig 2: System Architecture

B. WEKA AND ITS FUNCTIONS

WEKA is an open source data mining software which comes under the license of GNU. It was developed by the University of Waikato in New Zealand. This mining software can be freely downloadable. It was developed by using object oriented programing language. WEKA is mainly used in education and research purpose. Many tools and algorithms that support machine learning are available in it. It has a user-friendly interface so, anyone can use it easily. It has the option to check the programs written by us. WEKA also gives the platform that compares the different type's classification algorithms..

C. CLASSIFICATION ALGORITHMS HAVE BEEN EXAMINED.:

1. Naive Bayes
2. Logistic Regression
3. Random Forest
4. Support Vector Machine (SVM)

Naive Bayes

Naive Bayes is a classification technique based on Bayes Theorem [2] which provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$\underbrace{P(c\mid x)}_{\text{Posterior Probability}}=\frac{\overbrace{P(x\mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Logistic Regression

Logistic Regression is used when the dependent variable (target) is binary (0 or 1) [3][6].For example, to predict whether the attack is fack (1) or not (0). Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$=\log\left(\frac{p(y=1)}{1-(p=1)}\right)=\beta_0+\beta_1 x_2+\beta_2\cdot x_2+\_+\beta_p\cdot x_n$$

Random Forest

Random forests are a collective learning method for classification and regression [4][5] which creates a forest and makes it somehow random. Main advantage of random forest algorithm is, that it can be used for both classification and regression problems.

Support Vector Machine

Support-Vector Machines (SVM) [1] are supervised learning algorithm used for both classification and regression analysis which plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM is a frontier which best segregates the two classes (hyper-plane/ line). Fig 3 shows the sample hyper-plane of SVM.

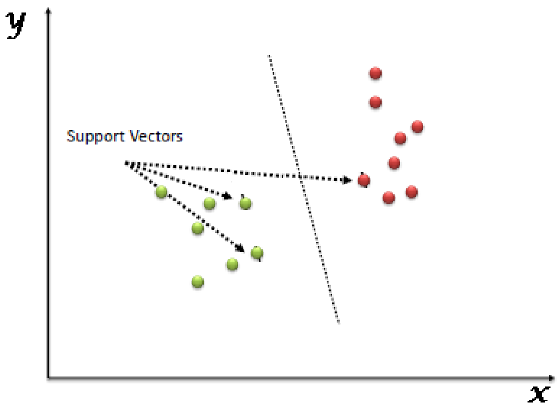

Fig 3 Support-vector hyper-plane

D. PARAMETERS USED

- TP = true positives: predicted instances are positive which are actually positive.
- FP = false positives: predicted instances are positive which are actually negative.
- TN = true negatives: predicted instances are negative which are actually negative.
- FN = false negatives: predicted instances are negative that are actually positive.
- Precision - % of selected items that are correct and are calculated as Precision (P) = TP / (TP+FP).
- Recall - % of correct items that are selected and the calculation for it is Recall (R) = TP / (TP+FN).
- F-Measure (F) - the Harmonic mean of precision and recall, calculated as F=2*R*P/(R+P).

E. DATA SET

A Meteorological data is indispensable for weather prediction and water resource planning. In this research work, the data were collected from Indian Meteorological Department Website on a monthly basis between 2016 and 2017 of the New Delhi region. Monthly data were collected on a day to day basis and different parameters such as Temperature, Humidity, Windy, Wind Direction, and Wind Chillness were collected and stored in the data file with the .csv format. Weather forecasting dataset has 20496 number of instances and 7 attributes with the last one as a class attribute. Table3 shows the attributes and details of weather forecasting dataset.

Table1: Attributes and data description of Meteorological dataset

Attribute	**Description**
Date	Current date and time
Temperature Numeric	The temperature values in Fahrenheit.
Temperature Nominal	This Attribute indicates the weather being below 60F as cool and above 60F as hot.
Humidity Numeric	Humidity is normally expressed as a percentage. A higher percentage means that the air–water mixture is more humid.
Humidity Nominal	This Attribute indicates humidity level where above 80% is high value and below 80% is normal.
Windy	This Attribute explains the windy weather conditions.
Wind direction	Wind direction is the direction from which it originates.
wind chill	This Attribute explains the amount of chillness in the wind.
Class label	Sunny, Rainy, Overcast

III. RESULTS AND DISCUSSION

Results obtained this research are based on different test options: k-fold cross-validation.

A. *Prediction: k-fold validation*

This research has used the K-fold cross-validation (k=10) method. This method divides a dataset into 10 folds, 9 folds are used for training, and the final fold is for testing..

B. *Correctly and incorrectly classified incidents of each classifier method*

Correctly and incorrectly classified incidents of each classifier method Classification accuracy is the degree of correctness in classification. The degree of correctness can be evaluated using various classifiers for individual instances in the data set. Larger test set provides a good assessment of classifier accuracy.

TABLE 2: CLASSIFIERS ACCURACY ON THE DATASET BASED ON 10- FOLD CROSS VALIDATION

Classification Method	Correctly Classified Incidents	Incorrectly Classified Incidents
NaiveBayes	94.4129	5.5871
Logistic Regression	88.6622	11.3378
Random Forest	98	2
Support Vector Machine	95.8914	4.1086

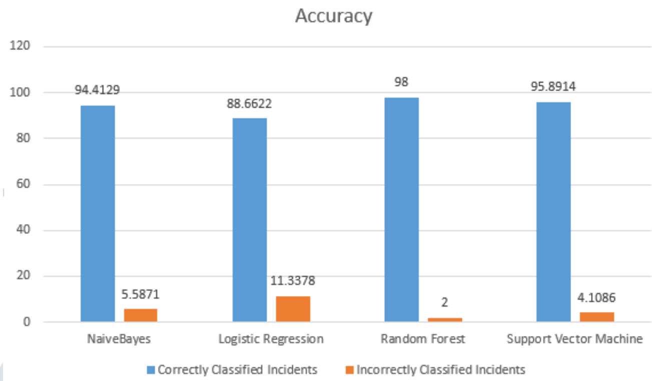

Fig: 4 Graphical view of accuracy for various classifiers

The random forest has identified a number of incidents correctly with 98%, followed by SVM having correct classification rate of 95.8% compared to other classifiers and Logistic Regression has determined least correct instances with 88.66%. Fig: 4 shows the graphical view of accuracy for various classifiers.

C. Performance measures calculated based on confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Confusion matrix contains some important parameters in order to calculate the performance of the classifications model. Figure 3 shows the weka confusion matrix. The analysis of Weather Forecasting Dataset on the basis of TP rate, FP rate, precision and F-Measure parameters are done. Table 3 shows the classification of testing data for different classes on TP rate, FP rate, Precision, Recall (R) and F-measure..

Table 3: Performance measures calculated based on confusion matrix

Classification Method	TP Rate	FP Rate	Precision (P)	Recall (R)	FMeasure (F)
NaiveBayes	0.944	0.020	0.954	0.944	0.946
Logistic Regression	0.887	0.094	0.892	0.887	0.888
Random Forest	0.98	0.02	1.000	1.000	1.000
Support Vector Machine	0.959	0.070	0.961	0.959	0.958

Table 3 shows the TP rate, FP rate, Precision, Recall and F-Measure, obtained by using the 10-fold cross-validation approach. Random Forest has the highest TP Rate (True Positive) by 0.98 and Recall values 100%, followed by SVM having TP rate by 0.95 and recall value of 95. Random Forest has greater precision and FMeasure when compared to other algorithms.

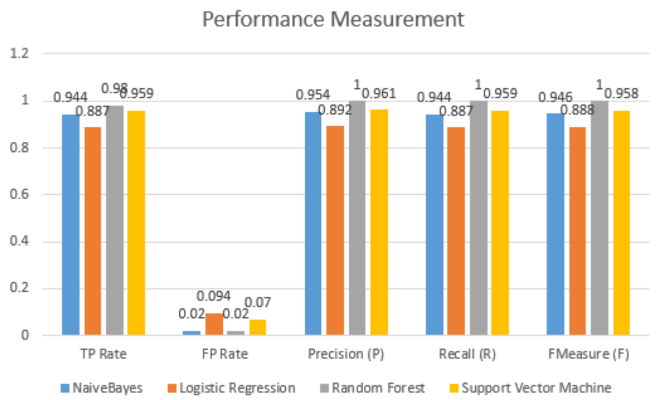

Figure: 5 Graphical view of performance measurement.

D. CLASSIFIER EXECUTION TIME AND ROOT MEAN SQUARE ERROR ON THE DATASET BASED ON 10-FOLD CROSS VALIDATION TEST MODE

Execution time is higher for SVM with 20.35 sec and j48 with 6.95 sec, while Forest time to build the model was the least with 0.09 sec, with NaiveBayes and Logistic Regression time for a model build is 0.12 sec and 1.55 sec, respectively. According to our experiment SVM consume higher time of 20.35 sec to build a classification model.

Table 4: classifier execution time

Classification Method	Time to Build the Model (Seconds)	Root Mean Squared Error
NaiveBayes	0.12	0.1831
Logistic Regression	1.55	0.1798
Random Forest	0.9	0
Support Vector Machine	6.95	0.2885

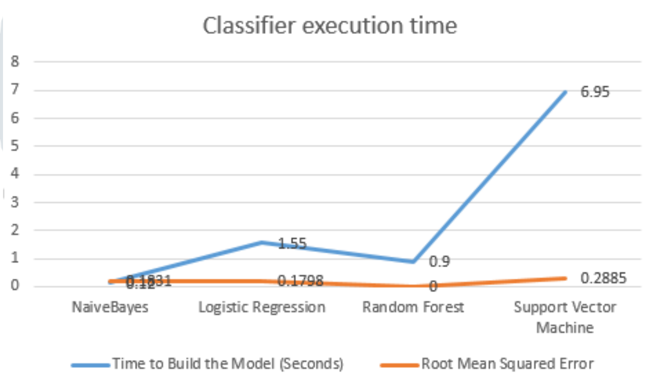

Figure: 6 Graphical representation of classifier execution time.

Conclusion: In this comparative study examined the performance of different classification algorithms. We use four important supervised learning classification algorithms in this experiment: Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). We have used Weather forecasting dataset to examine the performance of the classification algorithm. Based on the experimental analysis each algorithm gives a peculiar type of results. According to the experimental result, the random forest is a better algorithm because its execution time is low and the prediction rate is high. However, when the size and type of the data set changes there is the probability for change in the result. So the experimental results conclude that no classifier gives a perfect result. It necessitates the manufacture of a new classifier.

REFERENCES

[1]. Yujun Yang ; Jianping Li ; Yimei Yang, "The research of the fast SVM classifier method", IEEE Conference, 18-20 Dec. 2015.

[2] Haiyi Zhang and Di Li, "Naïve Bayes Text Classifier", 2007 IEEE International Conference on Granular Computing (GRC 2007)

[3] Yue Zhou and Jinyao Yan, “A Logistic Regression Based Approach for Software Test Management”, 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC).

[4] Simon Bernard ; Laurent Heutte and Sebastien Adam, “On the selection of decision trees in Random Forests”, 2009 International Joint Conference on Neural Networks.

[5] Angshuman Paul ; Dipti Prasad Mukherjee ; Prasun Das ; Abhinandan Gangopadhyay ; Appa Rao Chintha and Saurabh Kundu, “Improved Random Forest for Classification”, IEEE Transactions on Image Processing.

[6 Lei Liu, “Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning”, 2018 International Conference on Robots & Intelligent System (ICRIS).

[7] B.N. Lakshmi and G.H. Raghunandhan, “A conceptual overview of data mining”. 2011 National Conference on Innovations in Emerging Technology.

[8] Shivam Agarwal, ” Data Mining: Data Mining Concepts and Techniques”, 2013 International Conference on Machine Intelligence and Research Advancement.

[9] S. Umadevi ; K. S. Jeen Marseline, “A survey on data mining classification algorithms “,2017 International Conference on Signal Processing and Communication (ICSPC).

[10] Panigrahi Srikanth and Dharmaiah Deverapalli, “A Critical Study of Classification Algorithms Using Diabetes Diagnosis”, 2016 IEEE 6th International Conference on Advanced Computing (IACC).