

# A Comparative Analysis of Classification Algorithms on Weather **Dataset using Data Mining Tool**

# ATHUL C R<sup>1</sup>, Dr. Laksmi Mahesh<sup>2</sup>

<sup>1</sup>Scholar, SCMS, Cochin, Kerala, India <sup>2</sup>Assistant Professor, SCMS, Cochin, Kerala, India \_\_\_\_\_\*\*\*\_\_\_\_\_\_\*\*\*\_\_\_\_\_\_

**Abstract** - Data mining is the process of discovering insightful, interesting, and novel patterns, as well as understandable, descriptive, and predictive models from large-scale data. For analyzing the data which the software of Data mining to be used for analytical tools. There are different data mining tools are there, from these we selected one of the tools called Weka also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for analyzing the data and also predictive modeling. It involves algorithms and visualization tools that embed machine learning. The first phase called a training phase which analyzes the training data and the result of this analysis is a model that to make generalizations about how the attributes relate to the label. In the second phase called the testing phase at this time the model is applied to another, previously-unseen dataset where the labels are unknown. In a classification algorithm, the system tries to predict the label of each individual example. The Classification model tries to draw some output from the input values given for training. It will predict the class labels and categories for the new data. In our approach, we use classification algorithms like Decision Tree (148), oneR, Naive Bayes, and IBk. Thereupon we examine the competency of these four classification algorithms. This research work aims to compare the performance of a few data mining algorithms for predicting wind direction using historical weather data of New Delhi, India, which is collected from http://www.wundergrounds.com website.

# Key Words: Weather, comparison, machine learning, Weka, Prediction

# **1. INTRODUCTION**

Weather prediction is an important application in meteorology and has become one of the most scientifically and technologically challenging problem for meteorologists around the world. From the last few decades, the advancement and development in science and technology enable scientists to make better and precise weather prediction. The wind speed forecast in the wind energy sector is important for the following reasons like wind power system planning for unit commitment decision, load balancing decision, maintenance arrangement and energy storage capacity optimization. Wind speed is period arrangement information measured at various interims of time. Number of methods and techniques are used by the scientists to forecast rain; some of these techniques are more accurate than others.

Weather forecasting is the prediction of what the atmosphere will be like in a particular place by using technology and scientific knowledge to make weather observation. Long-established observations made at the surface of atmospheric pressure, temperature, wind direction, precipitation, humidity, etc are collected routinely from trained observers, automatic weather stations, or buoys. Now weather forecast has to rely on computer-based models that take many atmospheric factors into accounts. The input from the human is still required to select the best possible forecast model to base the forecast upon, which involves High-speed computers, wired and wireless sensors, metrological satellites, teleconnections, knowledge of model performance, etc. There is a huge amount of weather data available which is rich in information and can be used for weather prediction. Various data mining techniques are applied to the weather data to predict atmospheric parameters like temperature, wind speed, rainfall, meteorological pollution, etc. which tend to change from time to time and weather calculation varies with the geographical location along with its atmospheric parameters. Some commonly used data mining techniques for weather prediction are Decision Trees, Artificial Neural Networks (ANN), Naive Bayes Networks, Fuzzy Logic, Rule-based Techniques which includes Memory-based reasoning Techniques, and Genetic Algorithms.

# 2. METHODOLOGY

In data mining classification of large data set is a problem. Data mining has various techniques like classification, regression, clustering etc. This paper mainly focuses on the classification techniques having various algorithms which will help in classifying the records. The datasets contain instances or the classes and the attributes which helps in classifying the records. J48 Decision Tree, IBk, oneR and Naïve Bayes are the algorithms used for the analysis of the classification techniques. The research work mainly focuses on the comparative analysis of the classification algorithms which are Naïve Bayes, IBK, OneR and J48 on Weather forecast of wind direction dataset. The results of comparative analysis are anatomized to deduce best suited algorithm on the basis of definitiveness, execution time, correctly classified instances



and incorrectly classified instances Classification is a data mining technique and is a supervised learning having broad application. Classification technique classifies each item of a set into a predefined set of classes or groups. Among all the techniques in the data mining the apex technique is classification. Dataset is being inspected by classification and each instances of the dataset is considered. Checking the instances and considered by the technique are appointed class such that there will be less error in the model. Models defining the influential data classes inlying in a particular dataset are withdrawn using classification technique. The two states of the classification include application of the algorithm to construct the model and afterwards constructed model is tested contrary to an already defined dataset to measure the performance and definitiveness(accuracy) of the model. In this research work we have analyzed Naïve Bayes, OneR, J48 and IBk algorithms on Weather forecast of wind direction dataset.

Finally, it chooses the attribute that offers rules with minimum error and constructs the final decision tree

#### 2.1 Weka

WEKA known as Waikato Environment for Knowledge Analysis which is constructed in New Zealand in the University of Waikato. This machine learning software is written in Java. WEKA is a collection of visualization tools and algorithms for the predictive modeling. Different types of data mining algorithms can be tested using different type of datasets. The techniques which are supported by the WEKA are Data Processing, Classification, Clustering, Visualization Regression and Feature Selection. There are 5 interfaces in the tool and main user interface is explorer with which we work but all other interfaces provide same functionality just as the explorer.

## 2.2 Dataset used

## 2.2.1. Weather dataset

In this research work we have used Weather forecast of wind direction dataset. The main focus of this research is performance and evaluation of Naïve Bayes, IBk, J48, OneR algorithms. This dataset contains 83320 instances and 20 attributes. For analyzing the performance of the classification algorithms WEKA data mining tool is used.

- Number of attributes: 20
- sum of weights: 83320
- Total Size (in Mb): 7 Mb

G									Viewer								
Reit	tion new								_		_						
160	1: datetime_uk Nominal	2_conds	3dewptm Numeric	4: Jog Noverio	5: _hall I	E_heatindeem Numeric	7: _humio	8 _precipm 9 Numeria	. pressurem Nurvario	10rain Nomeric	11snow Numeric	12 Jempril Numeric	12 _Thunder Numeric	14: Jomado Numerio	15 visiti Numeria	15:_wdird Numets	17: _wgustm Numerio
1	255651	Clear	-20	5.0	50	29	75.0	20	1525.0	5.0	50	1.0	-50	5.0	1.0	20	20
2	25555155-1	Smake	20	5.0	5.0	2.0	87.0	2.0	-9999.0	5.0	5.0	2.0	50	5.0	5.6	5.0	2.0
3	25555157-6.	Unkno_	15.0	5.0	5.0	2.0	67.0	2.0	1517.0	5.0	5.0	2.0	5.0	5.0	2.0	5.0	2.0
£.	26555115-2.	18st	20	5.0	5.0	2.9	67.0	2.0	1519.0	5.0	5.0	2.0	5.0	5.0	2.8	295.0	20
5	25555117-5.	Unkno	27.0	5.0	5.0	2.0	67.0	2.0	1525.0	5.0	5.0	2.0	5.0	5.0	1.4	265.0	2.0
6	25555123-5.	Fog	2.0	1.0	5.0	2.0	87.0	20	1517.0	5.0	5.0	2.0	5.0	5.0	5.6	5.0	2.0
7	25555255-5	Light D.,	2.0	5.0	5.0	20	57.0	2.0	1558.0	5.0	5.0	2.0	50	5.0	20	125.0	20
8	25555257-1	Unkno_	2.0	5.0	5.0	2.0	67.0	2.0	-9999.0	5.0	5.0	2.0	5.0	5.0	5.1	5.0	20
9	25555213-2	Unkna_	2.0	5.0	5.0	2.0	67.0	2.0	-9999.0	5.0	5.0	2.0	5.0	5.0	1.1	5.0	20
10	25555214-5.	Linkno	2.0	5.0	5.0	20	87.0	2.0	-9999.0	5.0	5.0	2.0	50	5.0	1.5	50	20
11	26555214-5	Unkno_	2.0	5.0	5.0	2.0	67.0	2.0	-9999.0	5.0	5.0	2.0	5.0	5.0	1.0	5.0	20
12	25555214-6.	Unkno_	2.0	5.0	5.0	2.0	57.0	2.0	-9999.0	5.0	5.0	2.0	5.0	5.0	1.0	5.0	20
13	25555626-5.	Wides.	2.0	5.0	5.0	2.0	67.0	20	998.0	5.0	5.0	2.0	5.0	5.0	20	315.0	20
14	25555635-6.	Haze	2.0	5.0	5.0	2.0	67.0	2.0	1552.0	5.0	5.0	2.0	5.0	5.0	3.0	155.0	2.0
15	25555751-5	Haze	2.0	5.0	5.0	2.0	57.0	2.0	1552.0	5.0	5.0	2.0	5.0	5.0	3.0	125.0	20
15	25555723-6.	Drizzle	20	5.0	5.0	2.0	57.0	2.0	-9999.0	5.0	5.0	2.0	5.0	5.0	4.0	5.0	20
17	25555858-5.	Haza	24.0	5.0	5.0	2.0	67.0	20	1552.0	5.0	5.0	2.0	5.0	5.0	2.8	275.0	20
18	25555912-5	Haze	20	5.0	5.0	2.0	67.0	20	1556.0	50	5.0	2.0	50	5.0	3.5	295.0	20
19	25555914-1.	Haze	2.0	5.0	5.0	2.0	57.0	2.0	1553.0	50	50	2.0	5.0	5.0	35	35.0	2.0
20	25555928-5	Smake	2.0	5.0	5.0	2.0	67.0	2.0	1515.0	5.0	5.0	2.0	5.0	5.0	1.8	185.0	20
21	25551515-1.	Haza	2.0	5.0	5.0	2.9	67.0	20	1559.0	5.0	5.0	2.0	50	5.0	4.5	355.0	20
22	25551517-1.	Haze	2.0	5.0	5.0	2.0	67.0	2.0	-9999.0	5.0	5.0	2.0	50	5.0	4.0	50	2.0
23	25551518-5.	Haze	2.0	5.0	5.0	2.0	57.0	2.0	-9999.0	5.0	5.0	20	5.0	5.0	25	185.0	20
24	25551525-1.	Smoke	2.0	50	5.0	2.0	67.0	2.0	1558.0	5.0	5.0	2.0	5.0	5.0	12	45.0	20
25	25551523-1.	Smake	2.0	5.0	5.0	2.0	67.0	2.0	1557.0	5.0	5.0	2.0	5.0	5.0	1.0	145.0	20
25	25551152-2.	Mst	2.0	5.0	5.0	2.0	67.0	20	1515.0	5.0	5.0	2.0	5.0	5.0	1.0	50	20
27	25551116-1.	Unkno_	2.0	5.0	5.0	2.0	87.0	2.0	-9999.0	50	5.0	2.0	50	5.0	17	5.0	20
28	25551117-6	Unkno_	2.0	5.0	50	2.0	67.0	20	-99999.0	5.0	5.0	2.0	5.0	5.0	5.8	5.0	20
29	25551256-2	Smoke	2.0	5.0	5.0	2.9	67.0	2.0	-9999.0	50	5.0	2.0	50	5.0	5.6	50	20
30	25551216-1	Linkno.	2.0	5.0	5.0	2.0	67.0	2.0	.0999.0	5.0	5.0	2.0	5.0	5.0	5.6	85.0	20
														7	-		



International Research Journal of Engineering and Technology (IRJET)e-ISSVolume: 07 Issue: 05 | May 2020www.irjet.netp-ISS





# 3. **RESULTS**

This research work analyses different classification algorithms accomplishment for Weather forecast of wind direction dataset. Comparison of classifiers for Weather forecast of wind direction datasets done using criteria accuracy, correctly classified instances, incorrectly classified instances, error rate, and execution time to analyses the performance of the classification algorithms and its application domain is also discussed. Models for each algorithm are constructed using two methods mainly Cross-Validation with 10 folds out of which training set uses 9 folds and 1 fold for testing and Percentage Split in which 60% of the dataset is used for the training and 40% is used for the testing and output is given according to it. Figures are shown for the comparison of the different classifiers for Weather forecast of wind direction dataset using a 10-fold cross-validation testing bed. Applications are also discussed of these classifiers in the table. According to the table and research the execution time taken by the One Rule algorithm is least with 0.4 seconds followed by Naïve Bayes with 0.68 seconds, J48 algorithm with 3.28 seconds and Instance-Based Learner took much more time for execution which is 159.3 seconds. Accuracy of both J48 and One Rule classifier have 95.07%, Instance-Based Learner with 73.61%, and naïve Bayes with 43.18%. The naive Bayes classifier algorithm of accuracies have difference in comparison to others. Hence according to the data J48 and OneR algorithms are most accurate in case of 10-fold cross-validation method.

## 3.1. Final test accuracy of cross-validation

Analysis for Weather dataset of Wind Direction using percentage Cross-validation is done and this is as below. The final value represents the average accuracy of all the training and testing steps. This value is significant because it represents the accuracy or in other words rate of successful classification by the resulting model. The following figures from Figure 5.1 to 5.5 represents the final test accuracy of the various classification algorithms

## 3.1.1 J48

		V	Veka Expl	orer					- 0	•
Preprocess Classify Cluster Asso	clate Select attributes	Visualize								
Classifier										_
Choose NaiveBayes										
Test options	Classifier output									
O Use training set										1
O Supplied test set	Number of Le	aves : 2	10							1
Cross-validation Folds 10	Size of the	tree o 👘 d	19							
O Percentage split % 68										- 1
More options	Time taken to	build model	1 3.28 #	conda						- 1
	Stratifi	ed cross-vali	dation ++	-						- 1
Biom) write	Sumary									- 1
franci Tanta	Correctly Cl	estified Inst	ances	79208		95,0751	i.			- 1
Start Stop	Incorrectly	Classified In	stances	4103		4.9249				- 1
lesuit list (right click for options)	Kappa statis	sic		0.94	157					- 1
and the prime sector for oppositely	Nean absolute	e error		0.00	196					
00:15:38 - rules.OneR	Root mean sy	sared error		0.06	192					- 1
00:17:14 - trees J48	selecive ace	plute error	land (	8.95	105 1					
00:18:20 - bayes NaiveBayes	Total Durbar	a squared err	ar.	49.94	37.4					
a service and received and	Idnored Class	· Delense Tea	Cances		6					
	runnies ores	youkuppen, and	reatives							
	Detailed	Accuracy By	Class	i.						Ĩ
		TF Rate	FF Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	Cla
		0.956	0.000	1.000	0.956	0.977	0.974	0.998	0.994	Sor
		0.974	0.000	1.000	0.974	0.987	0.987	0.997	0.980	10156
		0.977	0.006	0.926	0.977	0.951	0.947	0.996	0.972	10.54
		0.979	0.000	1.000	0.979	0.989	0.988	0.998	0.990	Nes
		0.998	0.009	0.855	0.998	0.921	0.919	0.995	0.846	NSM
	-									10.00

Figure 3.2: Ttest Accuracy of cross-validation of J48

International Research Journal of Engineering and Technology (IRJET)e-ISSN: 2395-0056Volume: 07 Issue: 05 | May 2020www.irjet.netp-ISSN: 2395-0072

#### 3.1.2. Naivebayes

Preprocess Classify Cluster Asso	ciate Select attributes Visualize								
Classifier									_
Choose NalveBayes									
Test options	Classifier output								
O Use training set									
C Sutellad test sat									
0.000									
Cross-validation Folds 10	lime taken to sulld model	: 0.65 28	conda						
Percentage split % 66	Stratified cross-vali	dation	-						
Mess onlines	Summary								
white observer.	Commenter Changed and Taxa								
	Incorrectly Classified Inst	acces	333/0		43.1020				
Nom) _wdire	Rama statistic	0100700	0.36	2.5	Juine / A				
	Mean absolute error		0.07	13					
Start Stop	Root mean squared error		0.21	43					
and had felable affect for antiparts	Relative absolute error		66.67	82 8					
esuit list (right-click for options)	Root relative squared err	or	92.68	39 8					
00 15 39 - oller OneP	Total Number of Instances		83311						
00.13.36 - 10165.016R	Ignored Class Unknown Ins	tances		9					
00.17.14 - 0005 140									
UV-16/29 - Dayes Maived aves	=== Detailed Accuracy By	Class eee							
	IP Rate	TP Rate	Precision	Recall	F-Measure	HCC	ROC Area	FRC Area	CI
	0.305	0.100	0.355	0.305	0.328	0.218	0.595	0.270	No
	0.106	0.013	0.192	0.106	0.136	0.124	0.893	0.187	NB
	0.027	0.006	0.266	0.027	0.049	0.064	0.863	0.272	WN
	0.077	0.011	0.478	0.077	0.133	0.156	0.900	0.457	Ne
	0.287	0.013	0.556	0.287	0.378	0.377	0.965	0.536	W
	0.146	0.008	0.583	0.146	0.233	0.267	0.812	0.394	Ea
	0.799	0.029	0.610	0.799	0.692	0.679	0.975	0.778	ES
	0.598	0.014	0.623	0.598	0.610	0.595	0.950	0.656	SE
	4			1					115

Figure 3.3: Accuracy of cross-validation Naive Bayes.

# 3.1.3. Onerule

		1	Neka Expl	lorer:					-	
Preprocess Classify Cluster Assoc	ate Select attributes	Visualize								
Jassifier										-
Choose NaiveBayes										
fest options	Classifier output									
O Use training set	< 35	0.0 -> 3256								
C Supplied test set	< 36	0.0 -> WIN								- 1
	>= 3	60.0	> North							- 1
Cross-validation Folds 10	(79209/83311	instances of	orrect)							- 1
O Percentage split % 65										- 1
How extract	Time taken t	o build model	1 0.4 sec	ebnoc						- 1
More options										
	Stretifi	ed cross-val:	idation ==							
Nam) _wdire	· Summary									
	Correctly Cl.	assified Inst	Ances .	79206		95.0727				
Start Diop	Incorrectly	Classified In	stances	4105		4.9273	*			- 1
esult list (right click for options)	Happa statis	tic		0.94	157					
	Hean absolut	e error		0.00	50					
00.15.38 - rules OneR	Delative sha	olura error		5.47	0.0 8					
00:17:14 - trees.J48	Root relativ	s squared er:	TOT	32.93	66 8					
00:18:20 - bayes NaiveBayes	Total Number	of Instances		83311						
	Ignored Clas	s Unknown Ins	stances		9					1
	Detailed	Accurecy By	Class							
		TP Rate	FP Rate	Precision	Recall	F-Measure	HTT	BOC Lites	FRC Area	CIA
		0,956	0.000	1.000	0,956	0,977	0.974	0.978	0,962	Nor
		0.974	0.000	1.000	0.974	0.987	0.987	0.967	0.975	5256
		0.977	0.006	0.926	0.977	0.951	0.947	0.985	0.907	1036
		0.979	0.000	1.000	0.979	0.989	0.988	0.989	0.981	Ne.a
		0.998	0.009	0.854	0.995	0.921	0.919	0.994	0.853	REM
	4									7.8

Figure 3.4: Accuracy of cross-validation OneR.

# 3.1.4. Instance-based learner

Weka	Explorer					-	
e Select attributes Visualize							
ghboursearch LinearNNSearch -A l'weka c	re EuclideanDistar	ice -R first-l	ast."				
Classifier output							
Time taken to build model: 0. Stratified cross-validati Signary	1 seconds a						ŕ
Correctly Classified Instance	61333		73.6193	5			- 1
Incorrectly Classified Instan Rappa statistic Mean absolute error Root mean squared error Relative absolute error	es 21978 0.70 0.03 0.17 28.90	96 09 5 52 1	26.3807	,			
Total Number of Instances	83311	90 E					
Ignored Class Unknown Instanc							
Detailed Accuracy By Class							
TP Rese TP : 0.081 0.0 0.051 0.0 0.754 0.0 0.754 0.0 0.655 0.	ate         Preclains           0         0.643           0         0.712           0         0.779           1         0.630           2         0.703           1         0.507           6         0.507           6         0.650           0         0.731	Recall 0.638 0.410 0.706 0.799 0.649 0.630 0.789 0.697 0.520 0.600 0.781	F-Measure 0.841 0.630 0.709 0.639 0.623 0.788 0.700 0.514 0.624 0.794	MCC 0.013 0.619 0.606 0.761 0.619 0.592 0.776 0.608 0.608 0.608 0.618 0.767	ROC Area 0.809 0.009 0.049 0.049 0.090 0.024 0.002 0.051 0.770 0.051 0.770 0.003	PRC Ares 0.767 0.417 0.537 0.655 0.436 0.439 0.642 0.611 0.280 0.403 0.403	Cla Nor HNW Wor WOW Eas ESE SE SSE SSE SSE SSE
0.791 0.0	0.781	0,791	0,786	0.767	0.890	0.650	110
	Weis           g Bend athiosis         Mountain           g Bend athiosis <td< td=""><td>Web Tuplote           generations         Visuality           generations         Visuality           generations         Arreas use Excitacion           generations         Control of the second se</td><td>Uses Explore           Beneral atticutes         Winstein           Beneral atticutes         Winstein           Beneral atticutes         Ministein           Beneral atticutes         A Franks cone Excited and/Channes A Franks           Beneral atticutes         Scientific and/Channes A Franks           Beneral atticutes         Cone of the Science A Science A</td><td>Webs Digitizer           Bend athinks         Visuality           Description         Visuality           Description         Display           Consider any display         Afresta me build softly for the second softly fo</td><td>Web% Explore           Description         Visuality           Bend allows         Visuality           Bend allows         Visuality           Bend allows         Arway are Eurification Collarse of Residuality           Constraints Interpretation of Residuality         The scheme to Ruisid work? 16.101 seconds           end Residuality         Constraints           end Residuality         Scheme to Ruisid work? 16.101 seconds           end Residuality         Scheme to Ruisid work? 16.101 seconds           end Residuality         Scheme to Ruisid work? 16.101 seconds           Residuality         Scheme to Ruisid Residuality Rui</td><td>West Explore                0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows                 0 mean athlows in Woods               0 mean ath</td><td>Weak Explore         - 1           Beneral Millions         Weak Millions         - 1           Beneral Millions         Amounta         - 1           Beneral Millions         Amounta         - 1           Constraints         Lings of Millions         - 1          </td></td<>	Web Tuplote           generations         Visuality           generations         Visuality           generations         Arreas use Excitacion           generations         Control of the second se	Uses Explore           Beneral atticutes         Winstein           Beneral atticutes         Winstein           Beneral atticutes         Ministein           Beneral atticutes         A Franks cone Excited and/Channes A Franks           Beneral atticutes         Scientific and/Channes A Franks           Beneral atticutes         Cone of the Science A	Webs Digitizer           Bend athinks         Visuality           Description         Visuality           Description         Display           Consider any display         Afresta me build softly for the second softly fo	Web% Explore           Description         Visuality           Bend allows         Visuality           Bend allows         Visuality           Bend allows         Arway are Eurification Collarse of Residuality           Constraints Interpretation of Residuality         The scheme to Ruisid work? 16.101 seconds           end Residuality         Constraints           end Residuality         Scheme to Ruisid work? 16.101 seconds           end Residuality         Scheme to Ruisid work? 16.101 seconds           end Residuality         Scheme to Ruisid work? 16.101 seconds           Residuality         Scheme to Ruisid Residuality Rui	West Explore                0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows in Woods                 0 mean athlows in Woods               0 mean athlows                 0 mean athlows in Woods               0 mean ath	Weak Explore         - 1           Beneral Millions         Weak Millions         - 1           Beneral Millions         Amounta         - 1           Beneral Millions         Amounta         - 1           Constraints         Lings of Millions         - 1

Figure 3.5: Ttest Accuracy of cross-validation of IBK.

#### 3.1.5. Tabulated summary of results of cross-validation

	For Weathe	r dataset of Wind Directi	on	
Classifier	Naïve Bayes	OneR	IBK	J48
Testing Bed	Cross Validation	Cross Validation	Cross Validation	Cross validation
Applications	Text classification, Spam filtering, Online Application, Hybrid recommender system	Text document categorization by term association[32]	Random projection in dimensionality reduction: Applications to image and text data[17]	Emotion recognition, Verbal column pathologies.
Execution Time	0.68 seconds	0.4 seconds	0.01 seconds	3.28 seconds
Accuracy	43.18%	95_07%	73.61%	95.07%

Table 1: Comparison of classifiers for Weather dataset of Wind Direction using cross-validation testing bed



*Figure 3.6: Graphical representation of different algorithms accuracy and execution time using cross-validation method.* 

In the graph the abbreviation NB stands for Naïve Bayes, 1R for OneRule, IBK for Instance-Based Learner. The number of correctly classified instances in Naïve Bayes is 35976, One Rule with 79206, j48 with 79208 and instance-Based Learner with 61333. The incorrectly classified instances by Naïve Bayes is 47335, OneRule with 4105, j48 with 4103 and Instance-Based Learner with 21978.

## 3.1.6 Final test accuracy of percentage split

Now analysis for Weather dataset of Wind Direction using percentage split method is done. The final value represents the average accuracy of all the training and testing steps. This value is significant because it represents the accuracy or in other words rate of successful classification by the resulting model. The following figures from Figure 5.6 to 5.9 represents the final test accuracy of the various classification algorithms below.

#### 3.2.1. Naivebayes

Preprocess	Classify.	Cluster	Associate	Select attributes	Visualize								
lassifier													
Chosta													
Cuotes	nanebaje	-											
est options				Classifier output									
O Use traini	ing set												
O Supplied	test set	Det.		Time taken to	test model	on test s	plit: 1.68	seconda					- 1
C Cross-val	Edation Fi			Summary -									- 1
0.0			_										- 1
<ul> <li>rercenta;</li> </ul>	pe spit	78 00		Correctly Cla	ssified Inst	ances	14524		43.5842	4			
	More optio	15		incorrectly C	Lassified In	scances	18800		56.4158	•			- 1
				Hean shanlute	error		0.01	17					- 1
			-	Root mean act	ared error		0.23	38					- 1
iom) _wdire				Relative aboo	luce error		67.01	53 8					- 1
				Root relative	squared ers		\$2.46	04 8					- 1
Start		Site		Total Number	of Instances		33324						- 1
solt list (rinh)	Letick for a	etions)		Ignored Class	Unknown Ins	stances		4					- 1
		,	_										- 1
00/23-41 - ba	yes Natvei	ayes		Decailed	voontwol al	CLASS							- 1
					TP Rate	TP Bate	Precision	Recall	T-Measure	NCC	BOC Ares	FRC Ares	C14
					0.309	0.106	0.344	0.309	0.325	0.212	0.617	0.267	Nor
					0.116	0.012	0.225	0.116	0.153	0.144	0.900	0.215	NNW
					0.023	0.006	0.239	0.023	0.043	0.055	0.861	0.284	NIN
					0.079	0.010	0.523	0.079	0.137	0.169	0.912	0.505	Neo
					0.448	0.007	0.786	0.448	0,571	0.578	0.982	0.747	WEW
					0.144	0.010	0.540	0.144	0.228	0.255	0.802	0.366	Eap
					0.810	0.019	0.697	0.810	0.749	0.737	0.982	0.825	ESE
					0.772	0.022	0.588	0.772	0.665	0.659	0.950	0.797	32
					0.125	0.002	0.651	0.128	0.214	0.283	0.987	0.563	SSE
					0.346	0.005	0.601	0.346	0.439	0.448	0.968	0.482	NNE
					0.864	0.251	0.235	0.864	0.370	0.367	0.878	0.384	NW
				1.1				-					10.00

Figure 3.7: Accuracy of percentage split of Naive Bayes.



#### 3.2.2. J48

Preprocess Classify Cluster Associa	te Select attributes Visi	ualize								
lassifier										_
(August 100 0005 100										
Gibbse J48-C 0.25-M 2										
est options	Classifier output									
🔿 Use training set	luation on test sp	lit ===								
O Supplied test set Set	has an anal model		and the second							
Cross-validation Folds 10	KEI CO DESC HOUEL	ou veau s	giit: 0.1 1	counta						
Perrentane snlit % 60	nary									
	Ly Classified Inst	ances	31644		94,9586	\$				
Nore options	ctly Classified In	stances	1680		5.0414	*				
	tatistic		0.94	44						
and address	solute error		0.00	97						
iom)_wore	an squared error		0.04	99						
	e absolute error		9.03	84 8						
Start	lative squared err	ror	30.23	22 %						
sult list (right-click for options)	unber of Instances		33324							
	Class Unknown Ins	stances		4						
00:23:41 - bayes NaiveBayes										
00:24:21 - trees J48	alled Accuracy by	01855 ===								
	70 Date	TO Date	Descision	Devel1	-Verman	MOD	200 2man	00C 3res	(1) ere	
	0.954	0.000	1 000	0.954	0.977	0.971	1 999	0 991	Earth	
	0.972	0.000	1.000	0 972	0.986	0.996	0 997	0 976	1002	
	0,975	0.007	0.925	0.975	0.950	0.946	0.995	0.960	NW	
	0,978	0.000	0.999	0.978	0.988	0.987	0.998	0.986	West	
	0.999	0.009	0.855	0.999	0.921	0.920	0.995	0.854	WEW	
	0.850	0.012	0.852	0.850	0.851	0.839	0.983	0.860	East	
	0.993	0.000	0.999	0.993	0.996	0.996	0.999	0.994	ESE	
	0,980	0.000	1.000	0.980	0.990	0.990	0.998	0.984	SE	
	0.504	0.000	1.000	0.504	0.670	0.706	0.994	0.782	SSE	
	0.975	0.000	1.000	0.975	0.988	0.987	0.998	0.978	NNE	
	A				1.111					÷.

Figure 3.8: Accuracy of percentage split of j48.

# 3.2.3. ONERULE

Preprocess Classify Cluster Associa	te Select attributes	Visualize								
lassifier										_
Choose OneR -B 6										
est options	Classifier output									
O Use training set										
	Time taken to	build model	: 0.29 pe	conda						
U Supplied test set	Fusinatio	n on test ar	117							
Cross-validation Folds 10										
Descentions relit & 60	Time taken to	test model	on test s	plit: 0.06	seconda					
b i erceinalte abie in on										
More options	Summary -	-								
				01/14						
	Tacorrectly Cia	Dappified Indi	ences	1680		5 6414	5			
om) _wdire	Kanna statist	tie	a range b	0.94	44		÷			
	Hean absolute	error		0.00	159					
Start Stop	Root mean squ	ared error		0.07	17					
ult list (right-click for actions)	Relative abso	lute error		5.50	159 %					
	Root relative	squared ers	:or	33.35	F 900					
10:23:41 - bayes NalveBayes	Total Number	of Instances		33324						
10.24:21 - trees. J48	Ignored Class	Unknown Ins	tances		4					
10:25:22 - rules OneR			-							
	=== Detailed	Accuracy by	C1833 ====							
		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area	CI
		0.954	0.000	1,000	0.954	0.976	0.973	0.977	0.961	He
		0.972	0.000	1.000	0.972	0.986	0.986	0.986	0.973	10
		0.976	0.007	0.925	0.976	0.950	0.946	0.985	0.905	10
		0.978	0.000	1.000	0.978	0.989	0.987	0.989	0.980	ñe
		0.999	0.009	0.855	0.999	0.921	0.920	0.995	0.854	WS
		0.850	0.012	0.852	0.850	0.851	0.839	0.919	0.735	Ee
		0.993	0.000	0.999	0.993	0.996	0.996	0.996	0.993	ES
		0.980	0.000	1.000	0.980	0.990	0.990	0.990	0.981	SE
	4									10.0

Figure 3.9: Test Accuracy of percentage split of OneR..

#### 3.2.4. Instance-based learner



Figure 3.10:Test Accuracy of percentage split of IBK

	weather	ualaset of wir	id direction	
Classifier	Naïve Bayes	OneR	IBK	J48
Testing Bed	Percentage Split	Percentage Split	Percentage Split	Percentage Split
Execution Time	1.68 seconds	0.06 seconds	159.3 seconds	0.01 seconds
Accuracy	43.58%	94.95%	74.18%	94.95%

# 3.2.5. Tabulated summary of results of percentage split

Table 2: Comparison of classifiers for Weather dataset of Wind Direction using cross-validation testing bed

Comparison of classifiers for weather dataset of wind direction using percentage split method According to this test method that is percentage split it can be concluded that Naïve Bayes took 1.68s, Instance-Based Learner took 159.3s One Rule took 0.06s seconds and, j48 took 0.01s for execution. Accuracy of the J48 algorithm comes out to be 94.95% while same as that of One Rule, Naïve Bayes with 43.58% accurate and, Instance-Based Learner with 74.18% accurate. The number of correctly classified instances in Naïve Bayes is 14524, Instance-Based Learner with 24722, One Rule with 31644 and J48 with 31644. Number of incorrectly classified instances in Naïve Bayes is 18800, Instance-Based Learner with 5602, OneRule with 1680 and J48with 1680.



Figure 5.12: correctly and incorrectly classified instances in case of Percentage Split







Figure 5.11: Graphical representation of different algorithms accuracy in percentage split

## 4. DISCUSSIONS

Graphical representation of different algorithms accuracy in percentage split method. The abbreviations in the chart stands for Naïve Bayes, Instance-Based Learner, One Rule, j48.Graphical representation of correctly and incorrectly classified instances by the classifiers are:



*Figure 5.13: correctly and incorrectly classified instances in case of Cross-Validation* 

From the graphs it is analyzed that there is no such difference between the performance of the classification algorithms they have significant performances for the Weather dataset of Wind direction but on the basis of graph analysis J48 and One Rule classifiers is most accurate when using cross-validation method and percentage split and also Naive Bayes classifier has less accurate while compared to other classifier.

## 5. CONCLUSION

Comparison and investigations of the accomplishment of various classification algorithms is done using different criteria which are accuracy, execution time, correctly classified instances, incorrectly classified instances and error rate. In this work we carried out an experimental work to compare popular classification algorithms for wind direction prediction using various performance measures over weather data of New Delhi, India. The different measuring attributes play a pivotal role in giving precise weather forecast of wind prediction. According to the result evaluation it can be concluded that J48 and ONERULE are both have most accurate with 95.07% when 10 folds cross-validation method is applied for weather forecast of wind prediction dataset and for Percentage Split method J48, and ONERULE algorithms both have most accurate with 95.95% accuracy. In our case, J48, and ONERULE approach proves to be an efficient and acceptable method for wind direction prediction. Every algorithm has its advantages and limitations; it is difficult to choose the best algorithm. The prediction accuracy of the model can be increased by developing a hybrid prediction model where multiple machine learning algorithms are put to work together. For our weather dataset, it was concluded after analyzing various models of supervised learning that the J48, and OneR classification algorithms has appreciable level of accuracy and acceptance.

## **6. REFERENCES**

- 1) Olaiya, Folorunsho and Adesesan Barnabas Adeyemo, "Application of data mining techniques in weather prediction and climate change studies", International Journal of Information Engineering and Electronic Business, Vol. 4, No. 1, pp. 51, 2012.
- 2) Nikam, B. Valmik and B. B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach", Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on. IEEE, 2013.
- 3) A. Geetha, and G. M. Nasira, "Data mining for meteorological applications: Decision trees for modeling rainfall prediction", Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.



- 4) Ahmed, Bilal, "Predictive capacity of meteorological data: Will it rain tomorrow?", Science Information Conference (SAI), 2015, IEEE, 2015.
- 5) D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, 2001.
- 6) R. Mohammad, M. Butt Ahmed and M. Baba Zaman, "Predictive Analytics: An Application Perspective", International Journal of Computer Engineering and Applications, Vol. 9, No. 8, Aug. 2017
- 7) M. Kannan, S. Prabhakaran and P. Ramachandran, "Rainfall forecastingusing' data mining technique", 2010
- 8) Bingham, Ella; and Mannila, Heikki;"Random projection in Dimensionality reduction: applications to image and text data", in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001
- 9) R. Mohammad, M. Butt Ahmed and M. Baba Zaman, "Tools for Predictive Analytics: An Overview", International Journal of Scientific Research Engineering & Technology (IJSRET), Vol. 6, No. 7, pp. 748-750, 2017.
- 10) Dutta, Pinky Saikia and Hitesh Tahbilder, "Prediction of rainfall using data mining technique over Assam", IJCSE, Vol. 5, No. 2, 2014, pp. 85-90.
- 11) P.Langley, W.Iba and K. Thompson," An analysis of Bayesian
- 12) L.Breiman, "Random forests", Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001
- 13) Dubey and D. Akash, "Artificial neural network models for rainfall prediction in Pondicherry", International Journal of Computer Applications, Vol. 120, No. 3, 2015