

# Comparing various Machine Learning Techniques for Predicting the Salary Status

Suyash Srivastava, Deepanshu Sharma and Priyanka Sharma

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

February 10, 2020

# Comparing various Machine Learning Techniques for Predicting the Salary Status Suyash Srivastava\* and Deepanshu Sharma\*(Student)

# Ms. Priyanka Sharma\*(Assistant Professor)

\* SRM Institute of Science and Technology, NCR Campus, Ghaziabad.

#### Abstract

Supervised Learning and Unsupervised Learning method is used for classification of the data for predicting that which machine learning technique will classify the data sets of salary status of the people that who are less than or equal to 50000 salary or greater than 50000 salary more efficiently. We take the attributes as age, job\_type, ed\_type(education type), capital gain, capital loss, race, work hours per week, native country, salary status, relationship occupation, marital status, gender. We use four classifier methods Naïve Bayes, Random Tree, Random Forest, REPTree for classifying the data sets. After classifications we apply K-means algorithm for clustering the data.

#### Keywords

Supervised Learning ,Naïve Bayes ,Random Tree , Random Forest , REPTree ,K-means.

#### 1.INTRODUCTION

In this research paper, we use Weka Software for visualizing the data sets that we take for predicting the solutions for the data sets by taking the age attribute for taking predictions for the salary status.

#### WEKA

Weka is a software i.e. a group of machine learning procedures and proficiented to examine the data set by the Data Mining and Machine Learning techniques. The algorithms can be linked directly to the dataset. Weka software contains of data preprocessing, classification, regression, clustering, association rules, and visualization methods. It is also an appropriate approach for increasing new Machine Learning ideas. WEKA software is produced by the University of Waikato2. The Weka tool includes the four applications shown in Figure 1.



Figure 1: Weka Interface

Weka Tool

Weka Tool is used to find the outcome for the dataset. In  $2^{nd}$  diagram , encompasses the succeeding categories:

- Pre-process: Preprocess is used to ignore repeated data, missing data, error containing data and inconsistent data.
  - •Classify: It is used for the categorization task. A large quantity of classifiers are used as a part of Weka.

•Cluster: Grouping of the data sets into clusters.

•Associate: Create the connotation instructions for the data sets.

•Select attributes: Choose the attributes in the data.

•Visualize: 2-Dimensional design of the data3. Inside the paper 31,978 numbers of samples are used with machine learning algorithms such as Naïve Bayes,Random Forest,Random Tree and REPTree classifications and K-means used for clustering. It continues to analyze and find the important attribute for income datasets . The research paper is represented into four parts. Part 2 presents the Methods and Results &Analysis over income dataset implying Weka software as showed at Part 3. Part 4 concludes the paper.

Open IVe ] Open URL. ] Open DB ] Ger	undo	Edt.	Save
Rer			
Choose AllFilter			Apply Stop
arrent relation	Selected attribute		
Relation: income(1) Attributes: 13 Instances: 31978 Sum of weights: 31978	Name: SalStat Missing: 0 (0%) Di	Type: Nomini stinct: 2 Unique: 0 (0%)	×
diributes .	No. Label	Count Weight	
	1 less than or equal to 50,000	24283 24283.0	
New           1         Injap           2         Antipa           3         Antipa           4         Instructure           5         Occupit           6         Occupit           7         Antipa           8         Synchro           9         Synchro           10         Displayed           11         Displayed           12         Antraposet           13         Displayed           14         Displayed	Class: flatflat (Harro		Visualize A
Ranova		780	

Figure 2: Pre-processing Screen

# Visualising

Visualising all the attributes of the data set by plotting bar chart graph of all the attributes.



Figure 3: Bar Chart

# 2. METHODS

2.1. Classification method in Weka

It consists of three steps for data set in Weka.

- 1.Prepare the data sets.
- 2. Put a classification algorithm
- 3. Analyze the result.

First, organize the data sets in .csv extension and upload the data sets then, select classifying algorithms. At last, Analyze the results.

#### 2.1.1.Prepare the data sets-

Data set is used as a portion in Weka. It consists of various parts of attributes as age, JobType, EdType, maritalstatus, occupation, relationship, race, gender, capitalgain, capitalloss, hoursperweek, nativecountry and SalStat.

# 2.2. Classification Methods-

#### 2.2.1. Clustering

Clustering is an Unsupervised Machine Learning technique that involves the grouping of data points into clusters. If we given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.

#### 2.2.2. Naïve Bayes

Naïve Bayes is a kind of Supervised machine learning technique of mixture model that can be used for classification or for clustering (or a mix of both), depending on which labels for items are observed. ... Naive Bayes classification and clustering can be applied to any data with multinomial structure.

#### 2.2.3. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

#### 2.2.4. Random Tree

In mathematics and computer science, a random tree is a tree or arborescence that is formed by a stochastic process. Types of random trees include: Uniform spanning tree, a spanning tree of a given graph in which each different tree is equally likely to be selected.

# 2.2.5. REPTree

REPTree: algorithm is a fast decision tree learner it is also based on C4.5 algorithm and can produce

classification (discrete outcome) or regression trees (continuous outcome). This early random decision trees method combines bagging and random feature selection methods to generate multiple classifiers.

# 2.2.6. K-means Clustering

It aims to partition n-observations into k-clusters into which each observation belong to the cluster to the nearest mean, serving as a prototype of the cluster.

#### 3.ANALYZE THE RESULT

#### 3.1. Classifiers

Naïve Bayes, Random Tree, Random Forest, REPTree.

#### 3.2. Analysis of the Results

The result of the classification is analyzed and stated built on performance procedures. 10-Fold Cross-Validation technique is used to evaluate the implementation of assembling techniques. By this method, data is classified in ten equally weighted divisions , by the divisions 9 of those are utilised as exercise set and the last one is utilised as a testing set. This is used to elaborate the presentation of classifying techniques. Output is measured by Kappa Statistics, Mean Absolute Error and Root Mean Squared Error and ROC curve metrics7.

# 3.3. Kappa Statistics

The Kappa Statistics calculates the arrangement of prediction with the original class - 1 defines whole arrangement. If K is equal to 1 is perfect arrangement or If K is equal to 0 is chance of arrangement.

# 3.4. Mean Absolute Error

This calculates the real attributes explicitly whole magnitude of the individual errors. This is little smaller than the RMSE.

# 3.5. Root Mean Square Error (RMSE)

It analyzes the correctness. This finds the alterations in standards predicted by the model. The RMSE of the dataset will always be greater or equal to the MAE. If the RMSE = MAE, then every error are is of the equal magnitude.



Figure 4:Output of the Random Forest Classifier

Choose NaiveBayes		
st options	Classifier output	
Use training set Use training set Supplied test set Cross-validation Folds 10 Percentage spit % 05 More options  Sem) SalSitet Elast Diop	Transf Ware of Annual States, 1, 124 Warehouse Ware Transfilled States = 2005 Consenie (Canadina Damanese 2005 Disparation (Canadina Damanese 2005 Disparat	
sudi Had (nght-cick for options) 18:42:44 - traes.Bandmafreet 18:45:423 - traes.Bandmafree 18:45:422 - traes.Bandmafree 18:45:54 - bayes.HastwoRayes	and Detailed Rounder By Class see         Box DF Bass PF Bass	

Figure 5: Output of the Naïve Bayes Classifier

Choose NaiveBayes		
est options	Classifier output	
Use training set Supplied test set Cost-aulidation Folds 10 Percentage set More options. Nom SatStat	ener Routzled enzon-metadenten en- er Banary en Sameren (Lassellet Banarase 2017) 11.212 h Routzenergi Casalitation Banarase 2017 11.2122 h Routzenergi Casalitation Banarase 2017 11	
Staft         Stop           esuit list (right-click for options)         18:43:46           18:43:46         trees.RandomForest           18:43:46         trees.REFTree           18:49:22         trees.REFTree	Head Hommary By Class were           17 Hear 77 Hear 1 relation           6.64 0.447           6.74 0.447           6.74 0.447           6.74 0.447           6.74 0.447           6.74 0.447           6.74 0.478           6.74 0.478           6.74 0.478           6.74 0.478           6.74 0.478	
15:49:54 - bayes.BaiveBayes	Confusion Ratrix 8 b 0 Constrained as 1011 1702   4 - iser than or equal to 00,000 1123 4084   b + greater than 00,000	

Figure 6: Output of the Random Tree Classifier

Constanting of Destingend Destin	Choose NaiveBayers		
Unit Sector	lest options	Classifier output	
Section 2014         Section 2014<	Use training set     Supplied test set     Supplied test set     Coss-waldadox Fails     More agtions.      More agtions.      Statt     Stat	Transference Services Bandle 1 Transference 	
	esultist (ryht-chck for options) 21:43:44 - trees.landom/treest 21:43:423 - trees.landom/tree 21:43:23 - trees.landom/tree 21:43:24 - bayes.latveBayes	<ul> <li>Institute Researcy by Class we we have the first term of the first sector of the first sector</li></ul>	

Figure 7: Output of the REPTree Classifier

de hanne year bescher under bescher under under bescher under under bescher under under bescher under under bescher under under bescher under under under bescher under under under bescher under under bescher under under under under bescher under under bescher under under bescher under under under under under under bescher under under under under bescher under under under under bescher under under bescher under under bescher under under bescher under under under under under under bescher under under under under bescher under under under besch	Apple that is in the second	Choose Simple/Means - Init 0 - mae candidates	100-periodic-pruning 10000 -min-densit	(2.0-11-1.25-12-1.0-N 2-A "weka.core.Euclidea	nDistance -R firsHast"-1500 -num-sis	xts 1-8 10	
8 Unit Service	B to have been used         0.0191-0.1         0.0110-0.1           Provide status         0.0191-0.1         0.0101-0.1           Provide status         0.0191-0.1         0.0101-0.1         0.0101-0.1           Provide status         0.0101-0.1         0.0101-0.1         0.0101-0.1           Provide status         0.0101-0.1         0.0101-0.1         0.0101-0.1           Provide status         0.0101-0.1         0.0101-0.1         0.0101-0.1           Provide status         0.0101-0.0         0.0101-0.0         0.0101-0.0           Provide status         0.0101-0.0         0.0101-0.0         0.0101-0.0           Provide status         0.0101-0.0         0.0101-0.0         0.0101-0.0           Provide status         0.0101-0.0         0.0101-0.0         0.0101-0.0 </th <th>luster mode</th> <th>Clusterer output</th> <th></th> <th></th> <th></th> <th></th>	luster mode	Clusterer output				
Specific Market State         Sec.         5.1.10         1.1.10         1.1.10           Specific Market Scalable S	Bayes that is the second of the sec	Use training set		(81978.0)	(17785.0)	(14150.0)	
Prompetitive         Finance         Finace         Finance         Finance	International         Factors	O Supplied test set Set			45.5423		
Profile and ender the second of the	Manual procession         Engine methods         Engint dot Engind dot Engine methods         Engine methods		lege JohTume	St.STS Private	42.7953 Private	23-2967 Private	
Offense for device working         Description         Restrict-or-section         Restrict-or-section         Restrict-or-section           Start and section         main section         main section         main section         main section           section         main section         main section         main section         main section         main section           section         main section         main section         main section         main section         main section           section         main section         mai	) General device volume in transmission of tra	) Percentage split % 66	EdType	R5-strad	E5-grad	Some-college	
montpolicity         Tard speciality         Cathy speciality         Cathy speciality         Does - main           gene administ         main	menuncitie         menuncitie         ford-speciality         Cash - speciality         Cash - sp	Classes to clusters evaluation	maritelstatus	Married-civ-scouse	Married-civ-scouse	Sever-married	
Construction         minimum         minimum         minimum         minimum           general mathetic         minimum         minimum         minimum         minimum           general mathetic         minimum         minimum         minimum         minimum           general mathetic         minimum         minimum         minimum         minimum           Mathetic         minimum         minimum         minimum         minimum         minimum           Mathetic         minimum         minimum         minimum         minimum         minimum         minimum           Dial Of Impediations         minimum	B and states for searching         Battle bits         Battle	(Nom) SalStat	occupation	Prof-specialty	Craft-repair	Other-service	
sec         Back	unit         more         Raise         Raise         Raise           up         Raise         Raise	E Rose dustars to visualization	relationship	Rusband	Rusband	Not-in-femily	
upper efficiency         164 factor         154 factor         154 factor           But         Tome         154 factor         154 factor         154 factor           Distribution         Tome         154 factor         154 factor         154 factor           Distritin         Tome         154 factor	mer         100 Mail         100 Mail         100 Mail         100 Mail           me         mail         100 Mail         100 Mail         100 Mail         100 Mail           me         mail         100 Mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         100 Mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         mail         100 Mail         100 Mail         100 Mail         100 Mail           mail         mail         mail         mail         100 Mail         100 Mail         100 Mail		2000	White	White	White	
The dependence         0 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	The second sec		gender	Hale	Hale	Female	
Date         Description           Bit pred (able pred)         1.0.10           12/12         Superflower           Dial Structure         Lase that are speak to 0.0.00           Dial Structure         Lase that are speak to 0.0.00           Dial Structure         Lase that are speak to 0.0.00           The scenes to 0.0.01         Lase that are speak to 0.0.00           Lase that are speak to 0.0.01         Lase that are speak to 0.0.00           Lase that are speak to 0.0.01         Lase that are speak to 0.0.00           Lase that are speak to 0.0.00         Lase that are speak to 0.0.00           Lase that the speak to 0.0.00         Lase that are speak to 0.0.00           Lase that the speak to 0.0.00         Lase that are speak to 0.0.00           Lase that the speak to 0.0.00         Lase that are speak to 0.0.00           Lase that the speak to 0.0.00         Lase that the speak to 0.0.00           Lase that the speak to 0.0.00         Lase that the speak to 0.0.00           Lase that the speak to 0.0.00         Lase that the speak to 0.0.00           Lase that the speak to 0.0.00         Lase that the speak to 0.0.00           Lase that the speak to 0.0.00         Lase that the speak to 0.0.00	Implement         Open Control           Distribution         Distribution         Distribution         Distribution           Distribution         Distribution         Distribution         Distribution           The view to build work (full testing dec) / 6.25 seconds         Distribution         Distribution           The view to build work (full testing dec) / 6.25 seconds         Distribution         Distribution           The view to build work (full testing dec) / 6.25 seconds         Distribution         Distribution           - View 1 (htt)         Distribution         Distribution         Distribution	Ignore attributes	capitalgain	1064.3606	1569.0031	430.9995	
Date         Date <th< td=""><td>Date         Date         <thdate< th="">         Date         Date         <thd< td=""><td></td><td>capitalloss</td><td>86.7394</td><td>112.4104</td><td>54.5413</td><td></td></thd<></thdate<></td></th<>	Date         Date <thdate< th="">         Date         Date         <thd< td=""><td></td><td>capitalloss</td><td>86.7394</td><td>112.4104</td><td>54.5413</td><td></td></thd<></thdate<>		capitalloss	86.7394	112.4104	54.5413	
atter products for general Districe. See than or equal to 10,000 lease than or equal to 10,00	and any specific the Any specific 2023 On Encyclothers 2023 On Encyclothers There same to build small (field training any see Electronic factories Classical and realization of training any see Classical and realization and training any see	Start Stop	hoursperveer	40.4178	43.3505	36.743	
na na propositivity specific to inclusion of trade to inclusion of	name of performance and the second se	all first triable effects first mathematic	hativecountry	Dhited-States	Dhited-States	Dhited-States	
Differ Employablese           The waves to build hold (fb1) training decs) + 5.3 weaks           www.Real and employablese           Classese building decs)           2.778   100)           2.778   100)           2.778   100)	Diffe transition         The same to build model (full training decs) / 5.35 seconds           **** Read and realization on training set ==>           Classes           0           21705 1000           *	in the paper cards an optionally		year come of educe to action	year over of educt to action	year over or educt on action	
The same to build hold (full feating dees) : 5.3 seconds we find an evaluation on training set we Classrad backness 1 JUNE (44) 2 JUNE (44)	Ther space to build sold (full texaling mean) + 5.25 seconds we Hambi and reduction as training any sec Electronic Instances a	3:20.50 - SimpleHolleans					
The takes to Build Build (Hull Stating dec) + 0.35 seconds Build and emploiting in Stating ast Clustered Instances 0	Ther sales to huid such (full tending dee) / 5.25 events even Hub) and remaining on training etc Classes Training 0 1795 (940) 1 499 (149)						
The same to million work (find) tracking data y = 0.2 seconds 	The takes to half and will find training data) + 5.25 seconds were Halm and reministing on training per see Classical Instance 5 JTHS   540 4 JULY   5410						
<pre>www.Real and evaluation to training art === Classend measures 1 (2715 1 540) 1 (2115 1 640) 4 (2010 1 640)</pre>			Time taken to build model	(full training data) : 0.25 seconds			
<pre>*** Robi ad relation on training set *** Clustered hermone 0 TFF1 ( Sol) 1 All D ( Sol) 4 </pre>	er Bala af maintain a training es en g 2 2708 ( 3400 a 1000 ( 400						
Classes Instance 1 (1786) (40) 1 (100) (40)	Clastenet Instance 0 2778 ( 548) 1 3118 ( 549) 4		Hodel and evaluation	on training set			
1 1799 ( 90) 1 1799 ( 90) 1 1000 ( 00)							
0 27755 ( 546) 1 4 2639 ( 446)	a 1700 ( 190) ( 100 ( 00)		Clustered Instances				
			12245 / 5451				
			1 1/105 [ 504]				
	-				,		
	has a second		-)_[• <b>&lt;</b>				

Figure 8 Output of the K-means Clustering

#### 4. CONCLUSION

In this paper we compare unsupervised and supervised algorithms like Naïve Bayes, Random Forest, Random Tree REPTree and K-means for predicting the Salary Status of people using Machine Learning Technique . An income dataset contains 31978 original datasets are utilised. The dataset is analyzed by Weka software and the output of the classification is calculated in the form of their classification correctness by various measurements as Kappa Statistics, Mean Absolute Error and Root Mean Squared Error. It is observed that the correctness of REPTree Technique is most accurate in income dataset prediction in comparison of Naïve Bayes, Random Forest, and Random Tree classifier.

# References

- Dogra AK. A comparative study of selected classification algorithms of data mining. International Journal of Computer Science and Mobile Computing, IJCSMC. 2015 Jun; 4(6):220–9.
- D. Gao, YX Zhang. Random forest algorithm for classification of multiwavelength data, 2009 National Astronomical Observatories of Chinese Academy of Sciences and IOP Publishing Ltd.
- 3. Sushilkumar Kalmegh. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, Associate Professor, Department of Computer Science, Sant Gadge Baba Amravati University Amravati, Maharashtra- 444602, India.
- S.L. Ting, W.H. Ip, Albert H.C. Tsang. Is Naïve Bayes a Good Classifier for Document Classification?, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hum, Kowloon, Hong Kong jacky.ting@polyu.edu.hk.
- Hlaudi Daniel Masethe, Mosima Anna Masethe. Prediction of Heart Disease using Classification Algorithms, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

6. Weka Data Mining Tool. 23 Dec 2015. Available from: http://www.cs.waikato.ac.nz/ml/wek.