SALARY PREDICTION

ISCG7426 Data Mining Assignment 2

NOVEMBER 20, 2016

Song Lingyun - 1465073
Vahid Safinia - 1454841
Kritesh Thakur - 1466370

Contents

1.Introduction

The aim of this report is to predict whether the annual salary of a person could be more than $50k or less than $50k based on various factors like Age, Marital Status, Education, Occupation etc. With the demographic information of the population of the USA contained in this dataset, the analysis can be used for the salary evaluation and negotiation for people who are seeking a job, and can also be used in career planning or social investigation like who is earning the most.

A similar research has been done by the ITB on the same dataset. They have used Weka and Rapid Miner tools for their research. They have first done the exploratory and data analysis of the dataset. However, they removed a few attributes like country, marital status and education as there were a lot of duplicate and random values in those attributes. They worked on Naive Bayes, kNN and Rule Induction for the initial modelling. They divided the dataset in 3 types to test the same classifier on the different dataset type. There were slight variations in their results but all in all, Naive Bayes and NBTree were the best performers.

Several methods are used in this report targeting at the problems to be solved:

For Salary Prediction: machine learning methods such as Decision Trees, Logistic Regression and Support Vector Machines are used to train and build models based on the given dataset, and the salary can be predicted with this model for future data.

For Social Investigation: salary distribution on different criteria were analyzed such as age, education, race etc.

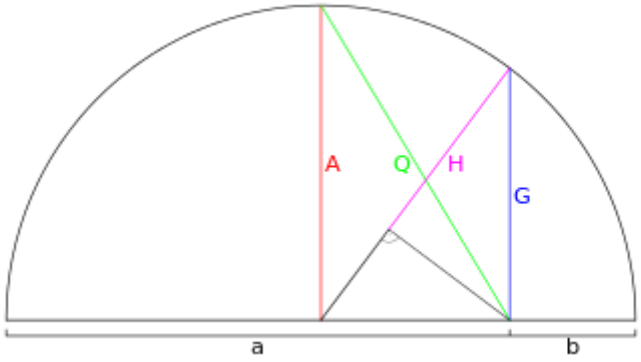
The evaluation criteria are based on the following metrics:

- Accuracy - it is the closeness of a measurement to the true value. It is calculated as:

$$\text{Accuracy} = \frac{True\ Positive(TP)+True\ False(TF)}{Total\ Instances}$$

- AUC – Area Under (ROC) Curve, is a measure of how well a parameter can distinguish between two classes, it will provide better verdict on classifier performance when the classes are imbalanced

- F-Measure – the harmonic mean of precision and recall, which conveys the balance level of precision and recall, only when precision = recall, F-measure get the biggest value. As shown in the below figure, the F-measure is **H**, **A** and **B** are recall and precision. You can increase one, but then the other decreases.



We are using the paired t-test in the end to compare the different classifiers and get the significant difference among them.

Work Structure

The workload was fairly divided among team members. Each member has used two classifier methods to solve the problem. The work structure of our team is shown in the table below:

Task	Person Responsible
Introduction & Criteria	Everyone
Data Exploration and Pre-processing	Everyone
Classification: J48, SVM	Song
Classification: NBTree, Logistic Regression	Vahid
Classification: RepTree, LMT	Kritesh
Evaluation and Comparison	Everyone
Conclusion	Everyone

2. Data Exploration

2.1 Dataset Description

The dataset was posted on “UCI Machine Learning Repository”, which was extracted by Barry Becker from the 1994 Census database. There are two datasets: train and test within the data folder, while with 32541 instances for train dataset, and 14532 instances for test dataset. To avoid overfitting, a validation dataset is necessary for the evaluation, thus both original train and test datasets were mixed and split again into Train, Validation, and Test datasets. Totally there are 48842 instances in the dataset, with 15 attributes, details as below:

age: The age of the observed sample, continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: The number of people the census takers believe that observation represents, continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: The education level, continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Gender, Female or Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

Salary: the class to be predicted, two possible outcomes, >50K or <=50K

2.2 Data pre-processing

Useless attributes, duplications and missing values were found in the dataset.

Useless attributes:

- Fnlwgt

  The number of people the census takers believe that observation represents, which is believed to be not important to the prediction

- Education:

  There is an equivalent attribute “education-num” which provide exactly the same information in the dataset

- Capital-loss

- Capital-gain

  Capital loss and capital gain are not considered in this dataset, because the purpose of this report is to predict salary instead of income.

All the left attributes were considered necessary for correctly predicting the

Duplications:

- 59 duplications are found in the whole dataset

Missing Values:

- There were missing values existing in “work class”, “Occupation” and “Native country”

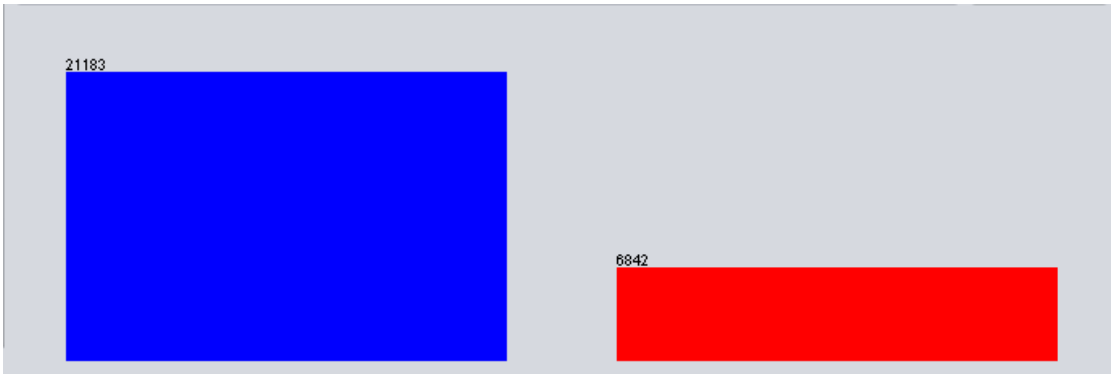
Data cleaning was done by Python in Jupyter, with the step of:

Remove missing values → remove duplications → remove useless attributes

The instances after data pre-processing remains 40037, with 11 attributes, whereby the last one “salary” as labelled class.

2.3 Data splitting

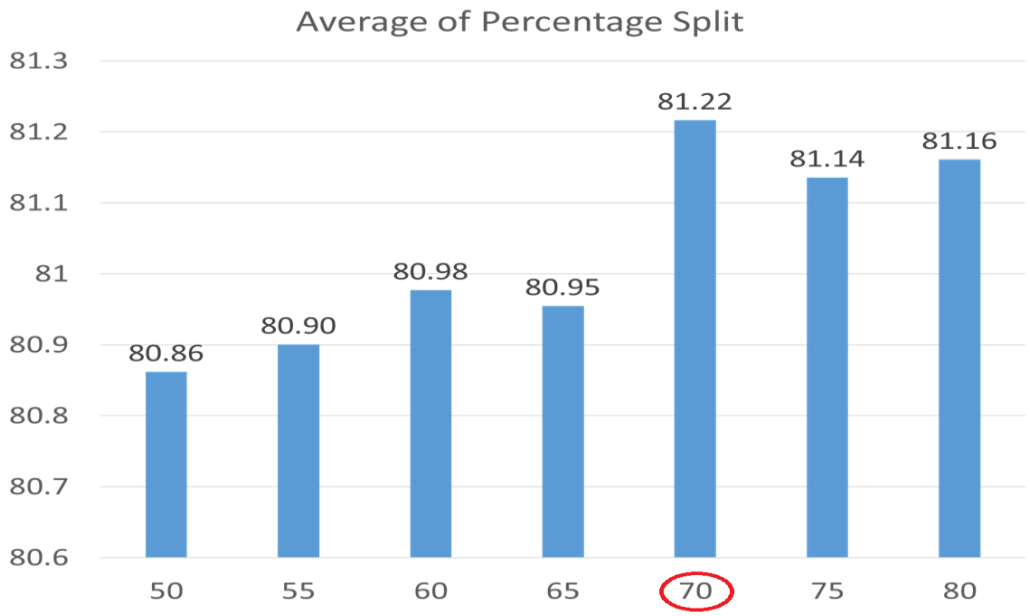
The class distribution after data cleaning is roughly 3:1 (<=50K:>50K), show as below:



Therefore classes are not imbalanced and no resampling methods (including over-sampling and under-sampling) were adopted.

The whole dataset was split into three sub-datasets: Train, Validate and Test. Several classifiers and split ratios were used to experiment the best split between train and test, with the summary table and statistical chart shown below:

Percentage	50	55	60	65	70	75	80	Average
Ibk	76.375	76.3262	76.2324	75.753	75.5236	75.4864	75.2728	75.85277143
RandomForest	79.3511	79.5221	79.532	79.028	78.9677	79.083	78.4278	79.13024286
Naive Bayes	79.4668	79.4929	79.6372	79.3961	79.7827	79.6824	79.5189	79.56814286
RepTree	81.0443	81.4804	81.6103	81.4843	82.1313	81.7962	82.3584	81.70074286
SVM	81.5911	81.6732	81.7286	82.3631	82.587	82.1222	82.4504	82.07365714
J48	82.0959	81.7725	82.0177	82.2129	82.5695	82.3536	82.5687	82.22725714
SMO	82.1432	82.0763	82.0835	82.318	82.4818	82.5744	82.4898	82.30957143
Logistic	82.1327	82.1289	82.1426	82.273	82.7009	82.7111	82.6344	82.38908571
LMT	82.1485	82.0997	82.1755	82.3856	82.6834	82.7216	82.9762	82.45578571
NBTree	82.2694	82.4327	82.6093	82.3331	82.736	82.8268	82.9105	82.58825714
Average	80.8618	80.90049	80.97691	80.95471	81.21639	81.13577	81.16079	

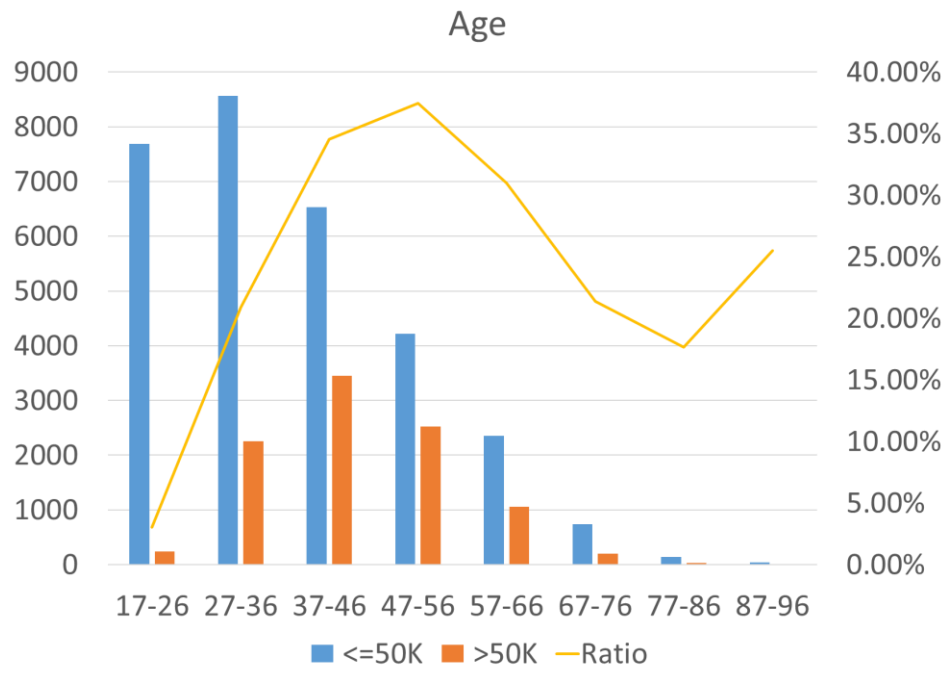


70:30 was found to be the best split ratio, while the 30% of testing dataset was split again into validate and test datasets. In conclusion, the data split ratio for this dataset was decided to be $\mathbf{Train}:\mathbf{Validate}:\mathbf{Test} = \mathbf{70}:\mathbf{15}:\mathbf{15}$

3. Data Analysis

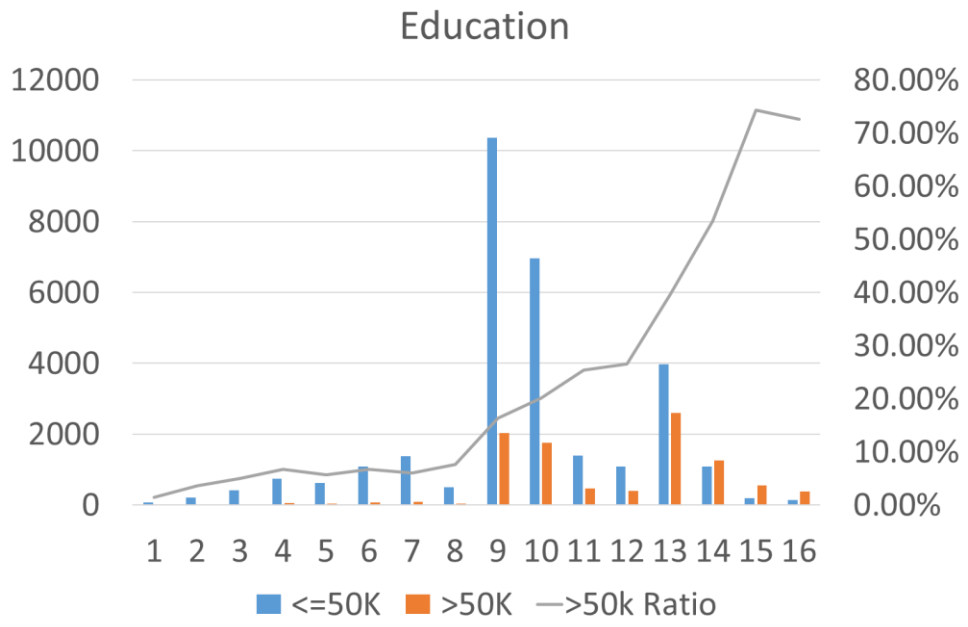
Some statistical analysis was done to better understand the dataset, while also provide more information for social investigations on the salary distribution with respect to different criteria.

1. Age



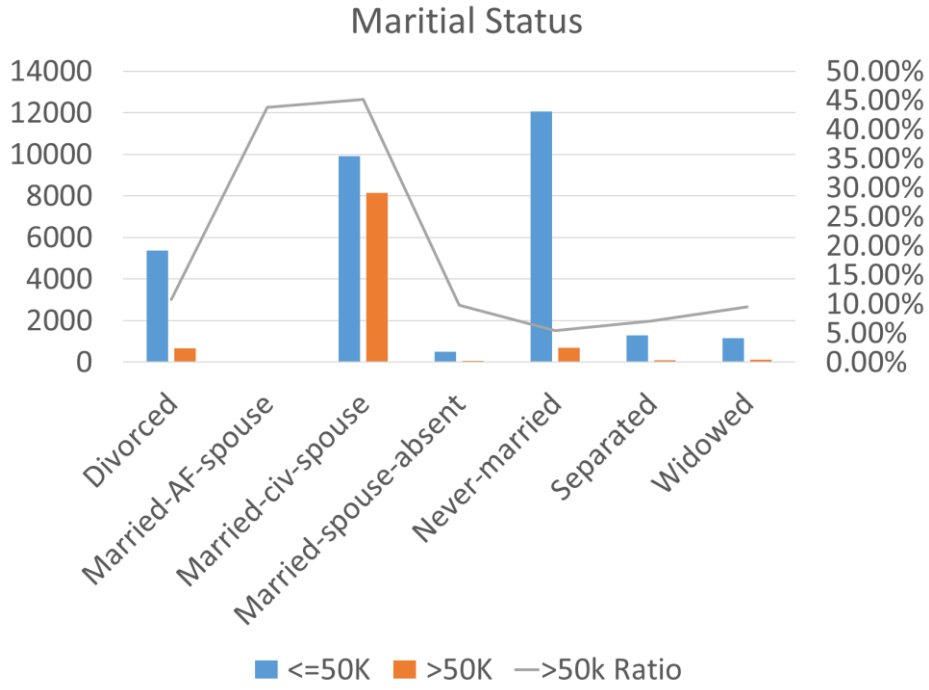
- The range of working ages in US is 17-90 (which is surprising)
- People aged from 47-56 tends to earn the most
- Most young people aged from 17-26 earn less than 50K

2. Education



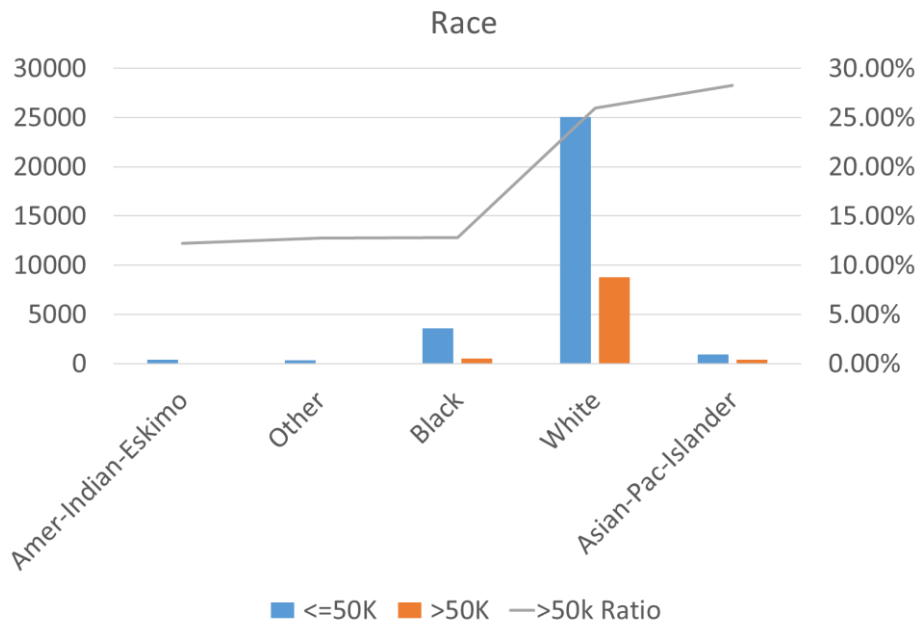
- The higher education level, the higher possibility one people can get better paid
- The possibility of earning more than 50K is increasing dramatically above level 13 (Bachelor)

3. Marital Status



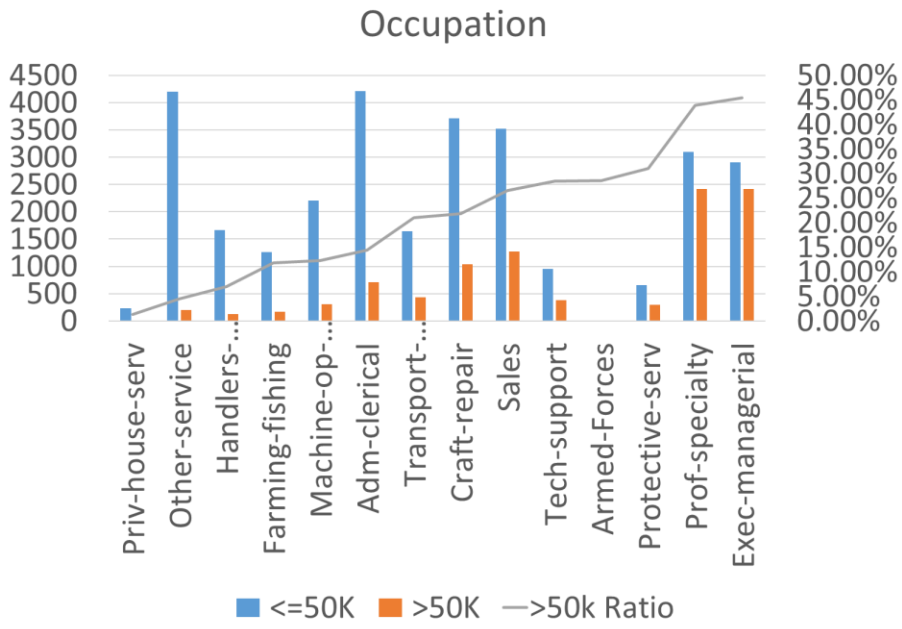
- Get married, then you will earn more!!!

4. Race



- The majority of working class is white people in US
- Although minority, Asians has the highest possibility to earn more

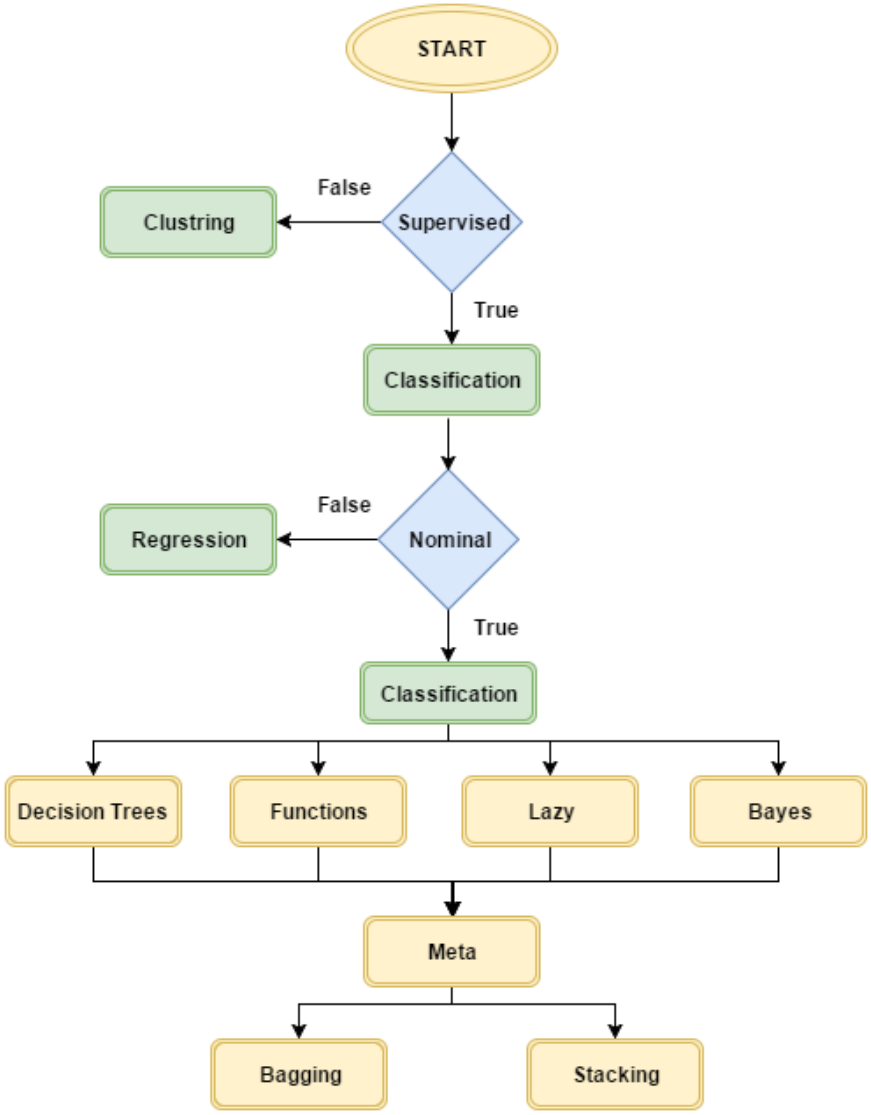
5. Occupation



- The best occupation for higher salary are: prof-specialty and exec-managerial
- The possibility for tech-support to earn over 50K is around 30%, moderate

4. Classifier Selection

The flow chart below shows how the classifiers were selected:

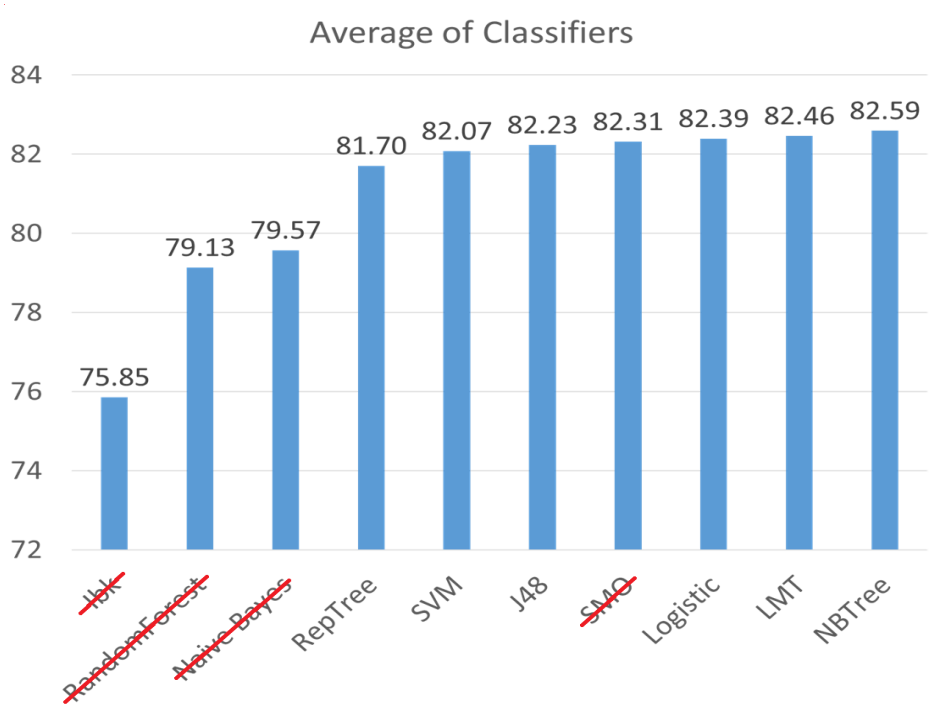


Almost all the categories of classifiers in Weka were tried out to find the best performance for the prediction, including classifiers of:

Category	**Classifiers Chosen**
Bayes	✓ Naïve Bayes
Functions	✓ SVM ✓ SMO

	✓ Logistic Regression
Lazy	✓ IBK
Decision Trees	✓ J48 ✓ REP Tree ✓ NB Tree ✓ LMT ✓ Random Forest

Experiments were done for all the classifiers above to test the performance on the dataset, with the statistical results below:



- IBK, Random Forest and Naïve Bayes excluded because of bad performance
- SMO excluded because it takes too long time to run, over 4 days running for parameter tuning, although with fair performance

Thus, the other 6 classifiers: SVM, J48, REP Tree, LMT, Logistic Regression and NB Tree are chosen in this report for the performance evaluation.

1. SVM

SVM (Support Vector Machines) is a classifier that uses support vectors to determine the best classification and separation among classes. Support vectors are the points that determine the boundary or edge of the classes. SVM is supervised machine learning method, which is suitable for the salary prediction dataset, and it applies certain kernel to arrange the data instances in such a way within the multi-dimensional space, to create a hyper-plane that separate the data instances into different classes.

In the salary dataset, all the nominal attributes are discretised, and transformed into numerical format, after which there will be 85 numerical attributes in the dataset. SVM build a 85-dimension kernel that separate the instances into two classes of >50K or <=50K.

Important parameters for SVM:

- C: c represents cost, which is the cost of misclassification. C trades off misclassification of training examples against simplicity of the decision surface, the surface of the kernel is "smoother" with a lower c value (fewer support vectors), while a higher c value gives the model more freedom to select more samples as the support vectors, which makes the kernel more complex.

- Gamma: gamma is a parameter used only in non-linear classification, it's the free parameter of the Gaussian radial basis function:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0$$

intuitively, the value gamma defines how far is the radius area of influence of one support vectors in training dataset. A higher value of gamma makes the influence of the support vector small or close, which leads to high bias and low variance, while a smaller value of gamma makes the influence of the support vector large, yielding low bias but high variance.

2. Logistic Regression

The goal of Logistic Regression is, given an observation X, and compute the probability of the class P(Y/X). It is called logistic because it uses special mathematical function to make sure that its predictions are truly probability and it is regression because the inputs for its function are features that multiply by coefficients like linear regression.

Logistic regression works very well with binary classes, which makes it good for the salary dataset.

3. J48

J48 is a classifier which implements the C4.5 algorithm. C4.5 builds a decision tree based on a set of training data using the concept of information entropy. At each node of the tree, C4.5 find one attribute that best split the instances into subsets based on information gain. J48 classifier is applicable to datasets with:

Class -- Binary class, Nominal class, Missing class values

Attributes -- Numeric attributes, Binary attributes, Missing values, Empty nominal attributes, Date attributes, Nominal attributes, Unary attributes

Decision tree is the most popular method in machine learning, with its advantage of being fast, and easy to interpret, while J48 is the most popular and basic methods among decision trees. With only 10 attributes in the dataset, J48 is expected to build a tree of moderate size. However, simple J48 tree tends to overfit as it may build a massive tree with too many splits that overfit the future data, that's why other error-pruning methods were experimented below.

Important parameters in J48:

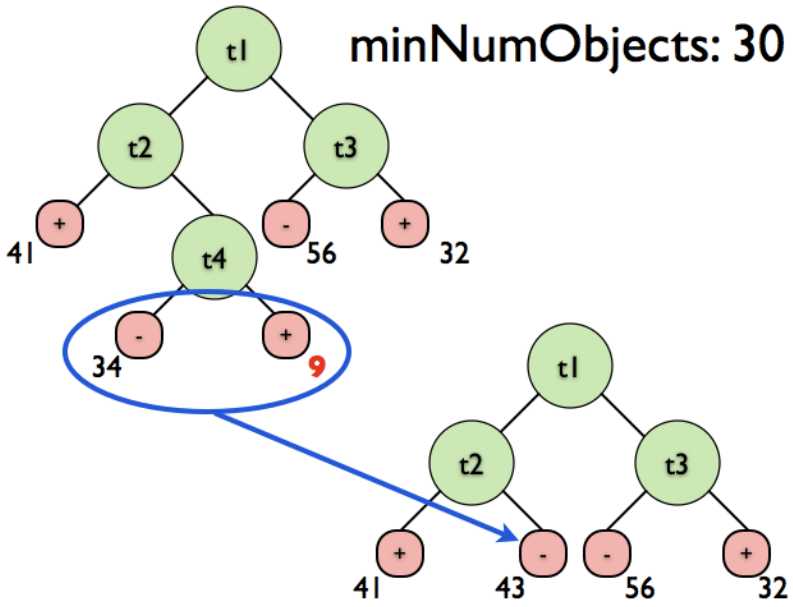
- Confidence Factor:

  Confidence factor indicates how confident you are in your training dataset during a post-pruning approach (back-fitting), with a lower confidence factor value, you are less confident with your training dataset, and the error estimate of each node is assumed to be high,

which increases the likelihood of this node to be pruned away during an error-pruning method.

- MinnumObj:

  Minimum number of objects is a factor used in pre-pruning, it limits the number of instances for each leaf in the tree, whenever a split is made which yields a child leaf that represents less than a minimum number of examples from the data set, the parent node and its children are compressed into a single node, thus reduce the number of leaves, and avoid over-fitting caused by a too detailed and complicated tree. An example of the influence of MinNumObj to the tree is shown as below [1]:



4. REP Tree

   REP Tree (Reduced Error Pruning Tree), same as J48 is also a fast decision tree learning method that build a decision tree with the training data, and predict the class label for new instances based on the tree. But different from J48, it has one more step of tree pruning using reduced-error-pruning method (back-fitting). After a tree has been built, each node is replaced with the most popular class starting at the leaves (back-fitting), if the prediction accuracy is improved, then the change is kept.

As J48 has been experimented in this study, REP tree is chosen to experiment the effect of reduced-error-pruning method on the dataset, to reduce error, avoid overfitting, and at the same time improve the simplicity of tree.

Important parameters in REP Tree:

- MaxDepth: MaxDepth puts a constraint on the depth (size) of the tree, making sure that all the leaves with depth > maxdepth specified will be pruned away, replacing them with the most likely class, thus reducing the complexity of the tree, and avoid over-fitting.

5. LMT

LMT (Logistic Model Tree) is another kind of decision tree which replace the leaves with regression functions. For a nominal attribute with k values, the node has k child nodes, and instances are sorted down one of the k branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise.

As both Logistic Regression and decision tree are experimented in this study, the combination of then may have a good result, which deserve a try.

Important parameters for LMT:

- Convertnominal: convert all nominal attributes to binary before building the tree, which means splits in the final tree will be binary
- Splitonresiduals: there are two possible splitting criteria for LMT, the default is to use C4.5 splitting criteria; the other tries to splitting to improve the purity in the residuals produces when fitting the Logistic Regression function

6. NB Tree

NB Tree finds a split for each attribute using J48. However, the choice of which attribute to use differs from what is done in J48. In NB Tree, 5-fold cross-validation is performed by applying Naive Bayes in each subset induced by the split, to estimate classification error. The attribute (i.e., J48 split) whose Naive Bayes models give the lowest classification error is chosen to implement the split for the current node.

In NB tree, to avoid too much splitting of the tree, splitting continues till reduction in error is greater than 5% and there are at least 30 instances in the node.

5. Evaluation

As mentioned above, three metrics: accuracy, AUC (Area under the ROC curve) and F1-measure are chosen for the evaluation criteria for this dataset. The importance of them are ranked as:

$$\text{Accuracy} > \text{F1} - \text{measure} > \text{AUC}$$

With the reason that:

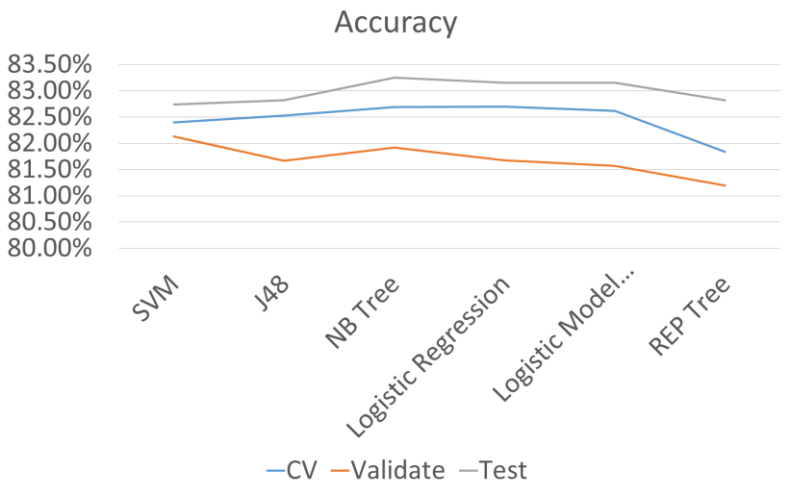
- No class is preferred more than the other, as the aim of this report is to predict one's salary level (range) as accurate as possible as per the criteria.
- When dealing with highly skewed dataset f-measure gives a more informative of an algorithm performance.

5.1 Baseline Evaluation

The baseline evaluation results are:

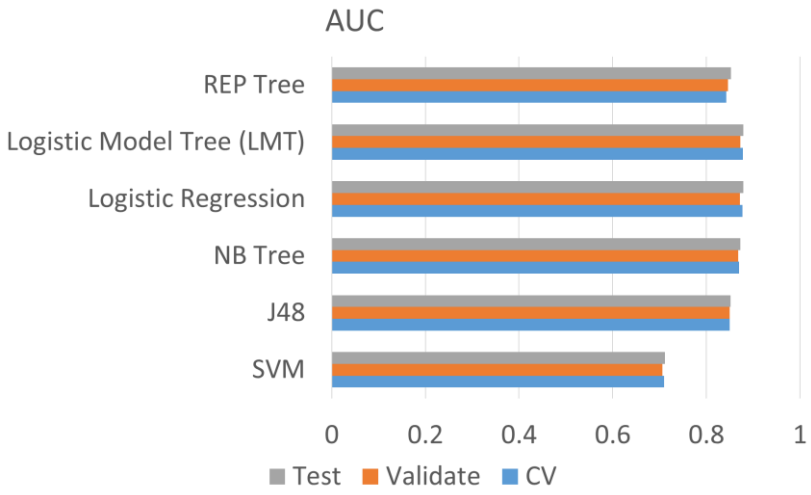
- Accuracy

Classifers	CV	Validate	Test
SVM	82.40%	82.13%	82.74%
J48	82.53%	81.67%	82.82%
NB Tree	82.69%	81.92%	83.25%
Logistic Regression	82.70%	81.68%	83.15%
Logistic Model Tree (LMT)	82.62%	81.57%	83.15%
REP Tree	81.84%	81.20%	82.82%



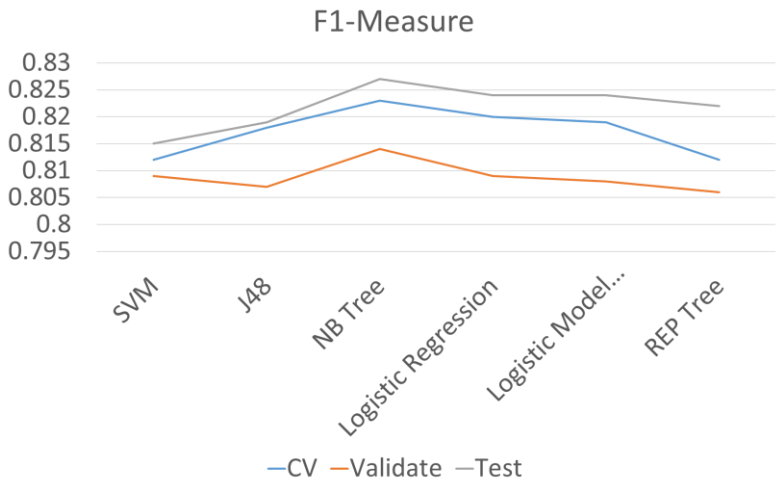
- AUC

Classifers	CV	Validate	Test
SVM	0.710	0.706	0.712
J48	0.850	0.850	0.852
NB Tree	0.870	0.868	0.873
Logistic Regression	0.877	0.872	0.879
Logistic Model Tree (LMT)	0.878	0.873	0.879
REP Tree	0.843	0.846	0.853



- F1-measure

Classifers	CV	Validate	Test
SVM	0.812	0.809	0.815
J48	0.818	0.807	0.819
NB Tree	0.823	0.814	0.827
Logistic Regression	0.820	0.809	0.824
Logistic Model Tree (LMT)	0.819	0.808	0.824
REP Tree	0.812	0.806	0.822



Conclusion: No significant difference found between the performance above, however, as NB Tree has the slightly better accuracy and F-1 measure performance, and the AUC is very close to the best, it is deemed as the best for baseline evaluation.

5.2 Parameter Tuning

Several methods were tried for different classifiers to find the parameters for their best results, as there are no important parameters in NB Tree and Logistic Regression that will influent their performance, parameter tuning is only tried on J48, REP Tree, SVM and LMT, using the following methods:

- Manually Tuning: J48, REP Tree
- Gridsearch in Weka: SVM
- Multisearch in Weka: J48, LMT

And the best parameter values found for the best performance of classifiers are:

Classifier	Parameters Tuned
J48	Confidence factor = 0.25, minNumObj = 8
SVM	cost = 100, gamma = 0.001
REP Tree	MaxDepth = 9
LMT	Split on residuals = True

And the classifier performance after parameter tuning are:

Accuracy	CV	Validate	Test
SVM	82.78%	82.15%	83.20%
J48	82.73%	81.95%	83.17%
Logistic Model Tree (LMT)	82.66%	82.38%	83.44%
REP Tree	82.10%	81.23%	82.60%

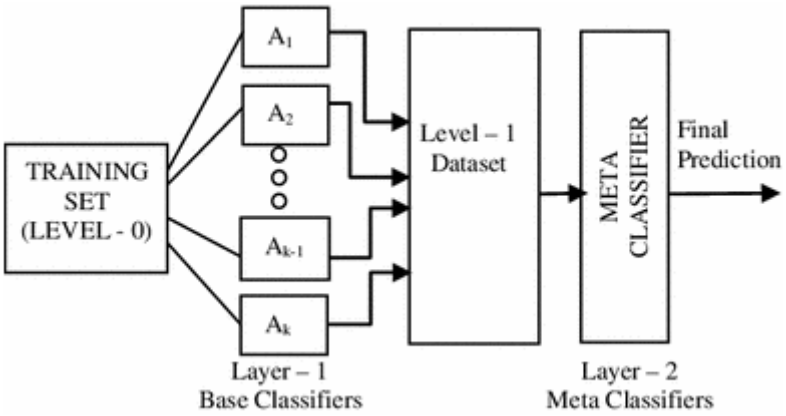
AUC	CV	Validate	Test
SVM	0.734	0.724	0.734
J48	0.856	0.859	0.855
Logistic Model Tree (LMT)	0.878	0.878	0.884
REP Tree	0.847	0.846	0.854

F-1 measure	CV	Validate	Test
SVM	0.821	0.814	0.824
J48	0.820	0.809	0.822
Logistic Model Tree (LMT)	0.821	0.818	0.828
REP Tree	0.814	0.805	0.819

Conclusion: No significant difference found between the performance above, however, as LMT has the slightly better accuracy, AUC and F-1 measure performance, it is deemed as the best for baseline evaluation.

5.3 Stacking

Stacking use different classifiers to learn on the same data, then instead of averaging the combined predictions, it uses meta-learners to generate a model based on the training set. By combining different classifiers, stacking correct some mistakes caused by a single classifier, the method is shown as figure below:



Two stacking methods were experimented, with the performances below:

Accuracy	CV	Validate	Test
NBTree & J48 & Logistic → LMT	82.8439%	82.4480%	83.2862%
SVM&Naïve Bayes&J48-->Jrip	83.9429%	81.7152%	83.0365%

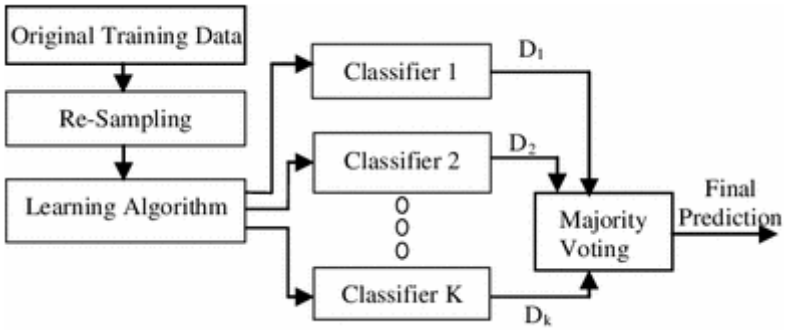
AUC	CV	Validate	Test
NBTree & J48 & Logistic → LMT	0.878	0.879	0.882
SVM&Naïve Bayes&J48-->Jrip	0.767	0.734	0.749

F-1 Measure	CV	Validate	Test
NBTree & J48 & Logistic → LMT	0.822	0.818	0.827
SVM&Naïve Bayes&J48-->Jrip	0.836	0.812	0.825

The first one: NB Tree & J48 & Logistic Regression → LMT (meta) was found to be the better one, with respect to all of the 3 criteria.

5.4 Bagging

Bagging creates K bootstrap samples of M size from training dataset, train the same classifier on the K samples, and combine the K resulting samples by simply majority vote. The method is shown as figure below:



Bagging works well on unstable classifiers, and can reduce the errors caused by a single unstable classifier. Thus, bagging method was only experimented on decision trees such as J48 and NB Tree, with the performances below:

Classifers	CV	Validate	Test
J48 Bagging	82.77%	82.05%	83.29%
NB Tree Bagging	82.88%	82.43%	83.60%

Classifers	CV	Validate	Test
J48 Bagging	0.866	0.869	0.864
NB Tree Bagging	0.875	0.879	0.880

Classifers	CV	Validate	Test
J48 Bagging	0.821	0.812	0.824
NB Tree Bagging	0.823	0.818	0.830

The performance of NB Tree bagging win out in all the three criteria, which makes it the better one in bagging method.

5.5 Conclusion

To find the best prediction performance for this dataset, all the best methods above are summarized and compared by t-test, with the conclusion below:

Accuracy	CV	Validate	Test
Baseline	82.69%	81.92%	83.25%
Parameter Tuning	82.66%	82.38%	83.44%
Stacking	82.84%	82.45%	83.29%
Bagging	82.88%	82.43%	83.60%

AUC	CV	Validate	Test
Baseline	0.870	0.868	0.873
Parameter Tuning	0.878	0.878	0.884
Stacking	0.878	0.879	0.882
Bagging	0.875	0.879	0.880

F-1 Measure	CV	Validate	Test
Baseline	0.823	0.814	0.827
Parameter Tuning	0.821	0.818	0.828
Stacking	0.822	0.818	0.827
Bagging	0.823	0.818	0.830

Conclusion:

- No significant difference found after t-test
- NB Tree bagging is the best method, with the best performance of Accuracy and F-1 measure, and a slightly worse performance on AUC.

The best performance for the salary prediction dataset is:

Classifier	Cross Validation	Validation	Test
Accuracy	82.88%	82.43%	83.60%
AUC	0.875	0.879	0.880
F1-measure	0.823	0.818	0.830

The reason that NB Tree Bagging performs the best in the salary prediction is:

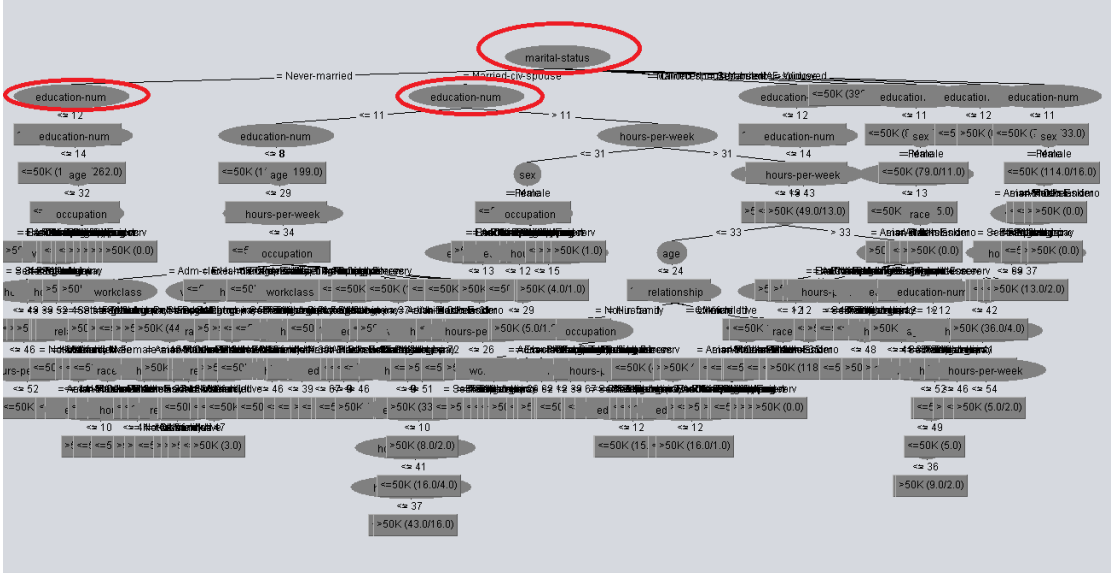
- NB Tree performs best because it is a hybrid approach suitable in learning scenarios when many attributes are likely to be relevant for a classification task, yet the attributes are not necessarily conditionally independent given the label. Moreover, the number of nodes induced by NB Tree is in many cases significantly smaller than that of J48.[2]
- Bagging method works well on unstable classifiers such as trees, and can reduce the errors caused by a single unstable classifier. Thus, NB Tree bagging

6. Discussion

6.1 Decision Trees

- J48 tree

The tree generated by J48 is shown below:



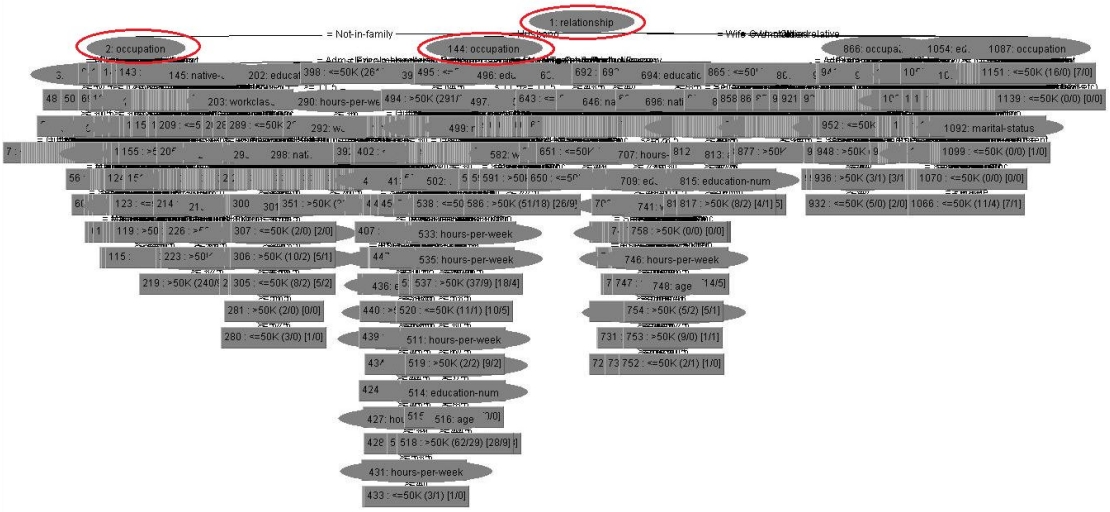

```
Number of Leaves  :     299

Size of the tree :      415
```

The root node is “marital-status”, while the attributes for the second layer is “education-num”, which means that “marital-status” contributes most information to the decision making of salary prediction, while “education-num” ranks the second.

- REP Tree

The tree after Reduce-Error-Pruning is:



Size of the tree: 1151

The root node is “relationship”, while the attributes for the second layer is “occupation”, which means that “relationship” contributes most information to the decision making of salary prediction, while “occupation” ranks the second.

- NB Tree

The Naive Bayes Tree is like:

Size of the tree: 138

The root node is “relationship”, while the attributes for the second layer is “occupation” “education-num” and “work-class”, which means that “relationship” contributes most information to the decision making of salary

prediction, while “occupation” “education-num” and “work-class” ranks the second.

- Logistic Model Tree

No tree generated by LMT, but the information can be concluded from the variable coefficient list below:

```
[age] * -0.01 +
[workclass=State-gov] * 0.13 +
[workclass=Self-emp-not-inc] * 0.17 +
[workclass=Private] * -0.03 +
[workclass=Federal-gov] * -0.22 +
[workclass=Local-gov] * 0.08 +
[workclass=Self-emp-inc] * -0.2 +
[education-num] * -0.14 +
[marital-status=Never-married] * 0.03 +
[marital-status=Married-civ-spouse] * -1.11 +
[marital-status=Divorced] * -0.1 +
[marital-status=Married-spouse-absent] * -0.11 +
[marital-status=Separated] * -0.14 +
[marital-status=Married-AF-spouse] * -1.15 +
[marital-status=Widowed] * -0.08 +
[occupation=Exec-managerial] * -0.36 +
[occupation=Handlers-cleaners] * 0.31 +
[occupation=Prof-specialty] * -0.3 +
[occupation=Other-service] * 0.41 +
[occupation=Sales] * -0.12 +
[occupation=Transport-moving] * 0.06 +
[occupation=Farming-fishing] * 0.44 +
[occupation=Machine-op-inspct] * 0.07 +
[occupation=Tech-support] * -0.25 +
[occupation=Protective-serv] * -0.28 +
[occupation=Priv-house-serv] * 0.48 +
[relationship=Not-in-family] * -0.16 +
```

Attributes will be transformed to binary in Logistic methods, with the highest coefficient value, “Marital-status” ranks the first to contribute most information for the classification.

- **Comparison:**

When all the attributes are ranked by Infogain or Gainrtio, the ranking results are:

- Infogain ranking

average merit	average rank	attribute
0.16 +- 0.001	1 +- 0	6 relationship
0.152 +- 0.001	2 +- 0	4 marital-status
0.086 +- 0.001	3 +- 0	3 education-num
0.084 +- 0.001	4 +- 0	5 occupation
0.08 +- 0.001	5 +- 0	1 age
0.052 +- 0.001	6 +- 0	9 hours-per-week
0.035 +- 0.001	7 +- 0	8 sex
0.019 +- 0	8 +- 0	2 workclass
0.01 +- 0	9 +- 0	10 native-country
0.009 +- 0	10 +- 0	7 race

- Gainratio ranking

average merit	average rank	attribute
0.081 +- 0.001	1 +- 0	4 marital-status
0.074 +- 0.001	2 +- 0	6 relationship
0.038 +- 0.001	3 +- 0	8 sex
0.033 +- 0	4 +- 0	3 education-num
0.028 +- 0.002	5.3 +- 0.46	1 age
0.026 +- 0	5.7 +- 0.46	9 hours-per-week
0.024 +- 0	7 +- 0	5 occupation
0.012 +- 0	8 +- 0	2 workclass
0.011 +- 0	9 +- 0	10 native-country
0.01 +- 0	10 +- 0	7 race

It is interesting that, no tree above is matching exactly the rankings of Infogain or Gainratio, however, it still can be concluded that the 4 most important attributes in Decision Tree classification methods for this dataset are: “Marital-status”, “relationship”, “education-num” and “occupation”.

The sizes of the trees are:

Tree	Size
J48	415
REP Tree	1151
NB Tree	138

REP Tree, as the reduce-error-pruning method, was supposed to reduce the tree size, making it less complicated and providing higher performance, however, as per the tree size and performance above, it is actually not working.

6.2 SVM and Logistic Regression

Linear SVM and Logistic Regression generally perform comparably in practice, but in this salary dataset, Logistic Regression is performing better than SVM, which means that the salary dataset is linearly separable.

6.3 Stacking and Bagging

- Stacking

As per the accuracy of stacking and the single classifiers below:

Accuracy	CV	Validate	Test
NBTree & J48 & Logistic → LMT	82.84%	82.45%	83.29%
J48	82.53%	81.67%	82.82%
NB Tree	82.69%	81.92%	83.25%
Logistic	82.70%	81.68%	83.15%
LMT	82.62%	81.57%	83.15%

The accuracy performance has been improved after stacking, which means that stacking worked. The reason is: stacking use different classifiers to learn on the same data, then instead of averaging the combined predictions, it uses meta-learners to generate a model based on the training set. By combining different classifiers, stacking correct some mistakes caused by a single classifier.

- Bagging

As per the accuracy of NB Tree bagging and the single classifiers below:

Accuracy	CV	Validate	Test
NB Tree	82.69%	81.92%	83.25%
NB Tree Bagging	82.88%	82.43%	83.60%

The accuracy performance has been improved after bagging, which means that bagging worked. The reason is: bagging works well on unstable classifiers such as decision trees, and can reduce the errors caused by a single unstable classifier. Furthermore, by bootstrap aggregating, the training samples will be increased (as there are overlapping instances in the K samples), then the classifier trains more and the performance will be improved.

6.4 Conclusion

The advantages and disadvantages of all the methods are listed below:

Classifier	Advantages	Disadvantages
J48	• Simple tree, easy to interpret and explain • Non-sensitive to linearly dependent features	• Sensitive to noisy data, biased • Interpretability goes down as number of splits increase
REP Tree	• Tree presented, easy to interpret and explain • Minimize the effect of incorrect or missing values in final representation of tree	• Performs bad at prediction of minority classes • More bias than J48
NB Tree	• Works well with big datasets	• Slower than decision trees and Naïve Bayes

	• Attributes should not be necessarily independent	
LMT	• The result is more intelligible than a committee of multiple trees or more opaque classifiers like kernel-based estimators	• Result is not as easy to interpret as a standard decision tree
SVM	• Works well with clear margin separation • Memory efficient as it uses a subset of training points in the decision function	• Becomes slow when the training dataset is large • Doesn't perform well when too much noise exists in the training dataset
Logistic Regression	• Easy to compute • Works well with binary classes	• Highly biased

7.Conclusion

This report aims to find the best method to predict salary according to different criteria such as age, marital-status, occupation, education etc., while at the same time, analysis on the dataset was performed to find the salary distribution within different criteria, and answer the question that: which criterion is the most important for a higher salary.

For the salary prediction, several classifiers together with parameter tuning and ensemble methods such as stacking and bagging are experimented, to find the best performance of prediction, while for the data analysis, statistical methods are adopted.

The best result of salary prediction comes from NB Tree Bagging, although not significantly better, it does win out other methodologies regarding to the three criteria of Accuracy, AUC and F1-measure, with the prediction accuracy of roughly 83%.

Lessons Learned:

- Two attributes: capital-loss and capital-gain were removed during the data pre-processing, because as stated above, the aim of this report is to predict salary, not income, the investment and capital issues are not cared thus not taken into account. However, during the classification process, when we tried out all the possible methods and ranked all the attributes in the original dataset, capital-loss and capital-gain turned out to have the highest infogain, which means they also contribute information to the salary prediction.
- Performance may differ when the same dataset is trained with the same classifier in different environment, for example, when the same salary dataset is trained by Logistic Regression in Python, the accuracy is 82.52%, while in Weka, the accuracy is 82.70%, because they use

different default values for parameters. So, for future work, different environment should be tried out to find the best result.

- Always try different methods. Some new methods such as LMT and NB Tree which is not very popular were tried out in this report, and NB Tree turned out to have the best performance. As each dataset has its unique characteristics, and classifiers has their advantages and disadvantages on different datasets, trying different methods help to find the best one.

8. References:

1- Drazin, S., & Montag, M. (n.d.). Decision Tree Analysis using Weka. Miami: University of Miami.

2- Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers:a Decision-Tree Hybrid. California,USA.

3- Kurokawa, M., Yokoyama, Y., & Sakurai, A. (2009, November). Averaged Naive Bayes Trees: A New Extension of AODE. Advances in Machine Learning, 191-205.

4- Landwehr, N., Hall, M., & Frank, E. (2003, September). Logistic Model Trees. Machine Learning: ECML 2003, 241-252.