**PROJECT REPORT**

**Classification of Real-Time Web Data using Analytics and Data Mining**



**INSY 5339 – PRINCIPLES OF BUSINESS DATA MINING**

### **GROUP 8: Harsh Vardhan Pande**

**Pranav Sathe**

**Samuel Geevarughese**

### **Surabhi Thakar**

**Uttara Krishnan**

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**TABLE OF CONTENTS**

1. DATA BACKGROUND................................................................................................................. 2

1.1 BACKDROP OF THE COMPANY........................................................................................... 2

1.2 WEB ANALYTICS FOR DIGITAL PLATFORM....................................................................... 2

1.3 DATASET & ATTRIBUTES..................................................................................................... 3

1.4 CLASS ATTRIBUTE................................................................................................................ 4

2. DATA CLEANING PROCESS...................................................................................................... 5

2.1 DATA CLEANING TOOLS...................................................................................................... 5

2.2 IRRELEVANT ATTRIBUTES.................................................................................................. 6

2.3 MISSING VALUES.................................................................................................................. 6

2.4 DERIVED ATTIBUTES............................................................................................................ 7

2.5 SKEWED DATA....................................................................................................................... 8

3. EXPERIMENT DESIGN................................................................................................................ 9

3.1 FALSE PREDICTORS............................................................................................................. 9

3.2 CLASSIFIER SELECTION....................................................................................................... 10

3.3 ATTRIBUTE SELECTION........................................................................................................ 10

3.4 FOUR CELL EXPERIMENT DESIGN...................................................................................... 11

4. EXPERIMENT RESULTS............................................................................................................. 12

4.1 RESULTS FOR EACH CLA SSIFIER...................................................................................... 12

4.2 SUMMARY OF RESULTS....................................................................................................... 16

5. ANALYSIS & CONCLUSION........................................................................................................ 18

5.1 ROC CURVE............................................................................................................................ 18

5.2 CLASSIFIER ANA LYSIS......................................................................................................... 20

5.3 ATTRIBUTE ANALYSIS........................................................................................................... 20

5.4 CONCLUSION......................................................................................................................... 20

6. REFERENCES.............................................................................................................................. 21

# 

# 

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**1. DATA BACKGROUND**

**1.1 OVERVIEW**

Across the rapidly growing domain of web, data is getting generated at a quick pace, which if used strategically for analysis can reflect a vast amount of hidden knowledge. Businesses are striving to make most of such data to increase their profit margins, potentially, to target maximum customers and to develop techniques through which new customers are introduced to their businesses in a web-based environment.

Currently, the information is growing at an exponential rate, organizations are dealing with the data torrent and are using analytics to increase their market value. The methods used to analyze data, which creates an expanding set of terms have been much in demand. In the industry of web analysis, emphasis is given on fixing the challenges of analyzing massive complex data sets when change is consistent with the data.

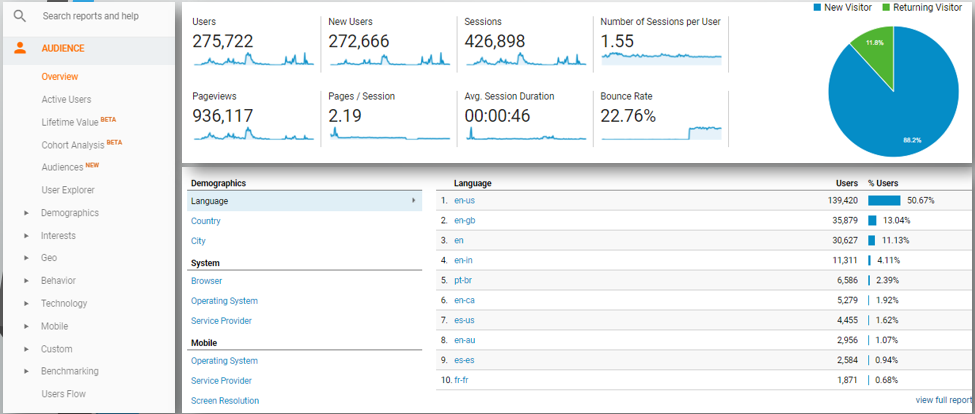
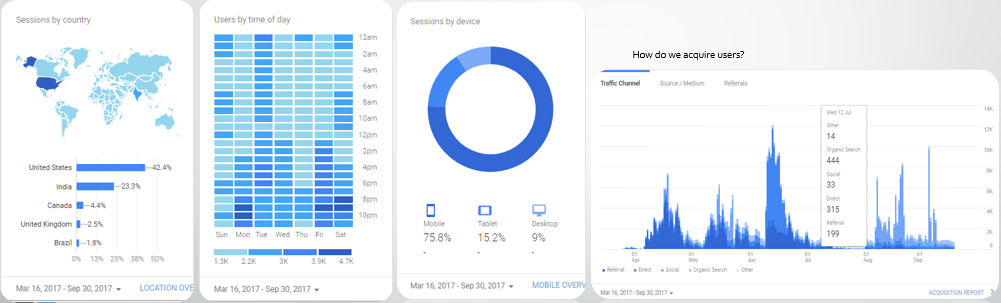


Figure 1

* 1. **BACKDROP OF THE COMPANY**

Our dataset comes from lifestyle Brand Company that deals with publishing articles on lifestyle at a digital platform. The website was launched in March 2017 and we have collected data from then up till six months.



*Figure 2*

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

* 1. **WEB ANALYTICS FOR DIGITAL PLATFORM**

The profit for the web domains is gained through the traffic that is coming through the site. Web analytics tells us how much traffic is entering and how it is affecting us. It helps us assess the favorite trends and helps in market research. The data collection and segmentation was done by the web analysis tool, i.e., Google Analytics.

Our dataset is based on behavior parameter taken from Google Analytics among others such as demographics, interests, geographic region, behavior, technology and mobile. The company growth is dependent on the revenue generation which is accomplished through page per clicks and advertisements which are dependent upon page views (under behavior→page views). Therefore, page views play an essential role in the analysis.

After gathering the data through Google web analytics, the collected data had a few discrepancies which were managed by Microsoft Excel through data cleaning purpose.

Once the cleaning and pre-processing was done the data is imported into the data mining tool, Weka.

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.

The question that we aim to answer in this paper is to determine the best classification algorithm to generate future predictions based on past data, to help improve/ increase the revenue of the company concerning the number of page views.

* 1. **DATASET AND ATTRIBUTES**

The data set that was provided contained ***13 attributes*** and approximately ***5000 instances***. The attributes are:

|  |  |
| --- | --- |
| Page URL | The pages visited, listed by URI. The URI is the portion of a page's URL following the domain name; for example, the URI portion of www.example.com/contact.html is /contact.html. |
| Page views | Page views is the total number of pages viewed. Repeated views of a single page are counted. |
| Unique Page Views | Unique pageviews are the number of sessions during which the specified page was viewed at least once. A unique pageview is counted for each *page URL + page Title* combination. |
| Avg. Time on Page | The average amount of time users spent viewing a specified page or screen, or set of pages or screens. |
| Entrances | Entrances is the number of times visitors entered your site through a specified page or set of pages. |
| Bounce Rate | The percentage of single-page sessions in which there was no interaction with the page. A bounced session has a duration of 0 seconds. |
| Exit % | %Exit is (number of exits) / (number of pageviews) for the page or set of pages. It indicates how often users exit from that page or set of pages when they view the page(s). |
| 1st Day | The total number of pageviews for each URL over the initial period of 24 hours over 1st day. |
| 10th Day | The total number of pageviews for each URL over the final period of 24 hours over 1oth day. |
| (1-10 days) total | Sum of the total number of pageviews for each URL over the initial period of 10 days, day 1 till the 10th day. |
| (11-20 days) total | Sum of the total number of pageviews for each URL over the period of next 10 days from 11th till the 20th day. |
| (1-10days) % change | Percentage change of page views over the initial period of 10 days, day 1 till the 10th day. |
| (11-20 days) % change | Percentage change of page views over the period of next 10 days from 11th till the 20th day. |

Table 1

* 1. **CLASS ATTRIBUTE**

Page views reflect the interest level of users and they also narrow down the chance for predicting future movements of user on a website. Thus, we chose our class attribute to be the **Percentage Change of Page Views in 11-20 Days**. As classification algorithms tend to identify certain characteristics so that it can group the class attribute. It generates a pattern having prior knowledge of the target attribute.

Our goal was to find out the best way to predict the page views. The quintessential data must be provided to a data mining algorithm to have information gain to make predictions for the class attribute. We derived four extra attributes namely 1-10 Days Total, 11-20 Days Total, 1-10 Days % Change as shown in Table 1 to bridge the gap between information provided by attributes and predictions of class attribute after running the algorithms.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

1. **DATA CLEANING PROCESS**

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a data set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty or coarse data.

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records).

As stated in section 1.2, we mentioned that our dataset comes from a real-world application. It is very likely that there might be reasons for the collected data to have noise or missing data inside it. Some of the most common reasons could be:

1. Data not available from the customer’s side.

2. Data entry error

3. Data migration and compiling errors

**2.1 DATA CLEANING TOOLS**

Prior to analysis on the data it goes through the process of cleaning and transformation. Data acquisition was carried out using Google Analytics which allowed us to import an excel sheet containing the raw dataset. This dataset was later transformed with help of various excel operations like replacing #DIV/0! with 0 and removing “ ”, ‘ ’ to fit into our criteria. Finally, Weka tool was used to decide if there were false predictors, to carry out discretization and for attribute ranking using attribute selection. This prepared dataset was then used for running various algorithms. The process that we follow is as shown in Figure 3.

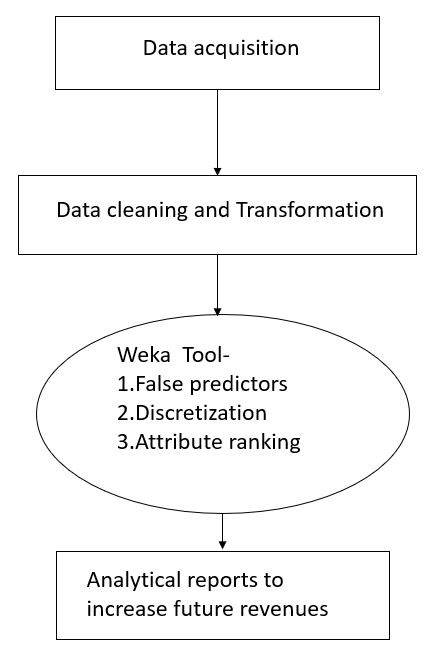
****

Figure 3

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

We utilized two tools for reconciling, normalizing and cleaning our data set:

### **a.** **Microsoft Excel**

### **b.** **Google Analytics**

Example of the use of Google Analytics:As the figure depicts the first step of our project was data acquisition which was completed by Google Analytics Technology. While users visit a web page/article google analytics at the back end keeps track of the entry point, number of page views, unique page views, time of visit. This all happens while user is completely unaware of the fact that his online behavior is getting noticed. This leads to acquisition of true realistic data. It also ensures that there are no null values present in the data. This results in authentic and disciplined data acquisition.

* 1. **IRRELEVANT ATTRIBUTES**

Irrelevant attributes can lead to misguiding the classifiers into building a correct model for predicting accuracy. Irrelevant attributes can be recognized with a few identification techniques:

a. They do not add any value to the class attribute in the dataset.

b. They convey the same level of detail as required by the analysis. Example- Both Zip and City name can be used to determine which city the example is from but if analysis needs details only at the city level then Zip becomes and irrelevant attribute.

Our data set had two such attributes:

**1st Day**- This attribute contained the views over 24 hours of the 1st day for a particular URL. This data was recursive and irrelevant as we had the details of 1st Day under 1- 10th Day Total available in our dataset.

**10th Day**- This attribute contained the views over 24 hours of the 10th day for a particular URL. This data was recursive and irrelevant as we had the details of 1st Day under 1- 10th Day Total available in our dataset.

**2.3 MISSING VALUES**

The original dataset didn’t have any missing values. It was because of efficient data collection used by Google Analytics by which we obtained our dataset.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

* 1. **DERIVED ATTRIBUTES**

As % change 11-20 days was out class attribute we needed to find attributes such as % change from 1-10 days, 1-10 days total and 11-20 days total. The values for these attributes were not readily available so we calculated them with the help of Excel.

Formulas used to calculate derived attributes,

1-10 total: Page views for day 1 to page views for day 10 total.

11-20 total: Page views for day 11 to page views for day 20 total.

1-10 % change- (|last day-first day| / first day) \* 100

11-20 % change- (|last day-first day| / first day) \* 100

The important point to be noticed out here is, our class attribute had values in terms of %change and therefore we needed % change from 1-10. The values in attribute 1-10 were the total number of views from day 1 to day 10. Similarly values in attribute in attribute 11-20 were the views from day 11 to day 20 which would give us enough information gain to the classifier to make prediction on class attribute.

* 1. **SKEWED DATA**

Real time environment is highly unpredictable in nature especially when it comes to dealing with behavioral traits of a user in his action of visiting a webpage. It solely depends on the user’s background, reason for visiting the page and how interesting does he find the content present on that page. In addition to this, there are many ways through which a website could be visited for example. a social networking site, a search engine or a straight visit. Different users carry different mindsets when they come on a website and therefore the data gathered to trace their online activities remains highly skewed in nature. Following are the cases which explain this scenario,

* A user may visit a web article of his interest once and not visit it again at all. On the other hand, a user may visit articles of same genre quite often with a same frequency. In a rare case, a user might visit multiple times and then he may not come at all on the same category of article or same articles.
* The entry points of users for visiting a web article could also vary which makes the algorithms job difficult since there may not be a pattern in visits of users on the webpage.
* Few users on the other hand are not users but they end up visiting a website accidently which is

the bounce rate of a web page.

In a broader perspective, we can say that the data could contain high variance due to difference in online behavior, difference in frequency of visits and the real-time nature of the project.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

1. **EXPERIMENTAL DESIGN**

**3.1 FALSE PREDICTORS**

Falsepredictors are directly related to target behavior. They define the events that happen at the same time or after the event relating to the target attribute. We have used One R to predict False predictors in our experiment.

In order to predict and eliminate false predictors we applied two methods:

1. **Domain Analysis**-In this method we analyzed the attributes that we used, to make predictions for the class attribute. We studied the domain from where these attributes were taken and after discussing attribute details with domain experts we concluded that few of the attributes could be removed from our dataset. As our class attribute was % change in page views, the total of day 1 through day 10 and day 11 through day 20 were irrelevant. Therefore, we decided to remove them.

Details of attributes that we eliminated:

a.1-10 days total- This attribute consisted attribute values which were the total of number of page views of day 1 to day 10.

b.11-20 days total- This attribute consisted attribute values which were the total of number of page views of day 11 to day 20.

**2.**  **Attribute accuracy using OneR**- After performing domain analysis we cross check our results with the OneR classifier. This algorithm tests for each attribute with the class attribute to check accuracy. If the accuracy result of this algorithm is coming close to 100 % then that means the dataset still has false predictors. OneR predicted one false predictor in our dataset.

Page URL- When we were running OneR we noted the prediction accuracy to be 99.37% and the prediction accuracy without page URL was 73.31%. So, clearly it was a false predictor and was removed from the dataset.

After performing domain analysis and OneR method we took off 3 attributes from our dataset namely 1-10 days total, 11-20 days total and Page URL. Initially we had 13 attributes after removing false predictors from our dataset we had 10 attributes in our dataset.

**3.2 CLASSIFIER SELECTION**

Once we remove the false predictors from our dataset we decided to select a classifier by which we can build a model. Our classifier selection process was solely based on two criteria’s:

a. High accuracy among all other predictors.

b. Reasonable stability.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

Discretizing the dataset and keeping ‘useEqualFrequency’ attribute as True with 3 bins we got the following results:

|  |  |  |
| --- | --- | --- |
| Classifier (66-33) % split | Accuracy | Accuracy (After SMOTE) |
| ZeroR | 31.7647% | 52.0527% |
| OneR | 51.8824% | 73.315% |
| Naïve Bayes | 51.7059% | 75.5821% |
| J48 | 51.7059% | 78.5821% |
| JRip | 48.7647 % | 78.5821% |

We selected the following classifiers:

**1.J48 (tree)** - It is a decision tree based algorithm which emphasizes on building a classification tree to prepare decision rules. It works mainly with DTL (Decision tree learning) process to find the most efficient attribute to increase the prediction accuracy. It is famous for building **high accuracy** models in the realm of data mining

**2.Naïve Bayes (Bayes)** – Naive is known for building a highly stable model which gives relatively less variance. It works on finding the probabilities to decide the attributes from dataset using which a model could be built. An advantage of Naive Bayes is that it requires small number training set to estimate parameters necessary for classification.

**3.JRip (meta)** – This is one of the basic and popular classifier. The incremental reduced error Ripper(JRip) proceeds by taking into consideration all the examples of a decision in training data as a class and then finding the rules that cover all the members of that class. Same process is repeated until all classes are covered.

**3.3 ATTRIBUTE SELECTION**

The attribute selection process is a part of the experiment design as one of the two factors in our experiment design is ‘**Number of Attributes**’ which is explained in the next section.

**ClassifierSubsetEval**: There are many methods to find the most valuable attributes using Weka. We picked the ClassifierSubsetEval method because it uses classifiers to evaluate attributes. Since, we have decided the classifiers we will be using in step 3.2, we can use those classifiers to determine which of our attributes are the most useful.

We selected the following three attributes as our ‘Selected Attributes’.

### **1.Age**

### **2.All Riders**

### **3.Annual Premium-** Discretizing Annual Premium attribute for all experiment runs gave us a better model with higher prediction accuracy.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**3.4 FOUR CELL EXPERIMENT DESIGN**

**Two Factor Design**: Our experiment design contained of two factors:

**1.** Factor 1 (F1):**Noise**

**2.** Factor 2 (F2): **Percentage Split**

F**our Criteria of the Design**: The two factors are to be divided up into 4 criteria by keeping one factor constant and varying the other factor between two values and vice versa. This is illustrated in the table blow.

a.**F11, C1**= Without Noise + Percentage Split of 20%/80%

b.**F12, C2**= Without Noise + Percentage Split of 80%/20%

c.**F21, C3**= With Noise + Percentage Split of 20%/80%

d.**F22, C2**= With Noise + Percentage Split of 80%/20%

With the above mentioned four criteria we are now ready with our experiment design. The selected attributes in the section 3.3 above would be a part of the selected attributes criteria.

**Note**: In order to make our training and test data truly representative as the data might lose its properties due to sampling and while running the classifiers, we are doing ten runs for each criterion, each classifier with a **distinct seed value**.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

# **4. EXPERIMENT RESULTS**

# **4.1 RESULT OF ALL CLASSIFIERS**

The table below describes the 12 possible combinations of our 4 criteria with the 3 selected classifiers. We ran each of these combinations 10 times and averaged their accuracy and variance:

|  |
| --- |
| **E1**= Performance of J48 when, Without Noise + Percentage Split of 20%:80% |
| **E2**= Performance of J48 when, Without Noise + Percentage Split of 80%:20% |
| **E3**= Performance of J48 when, With Noise + Percentage Split of 20%:80% |
| **E4**= Performance of J48 when, With Noise + Percentage Split of 80%:20% |
| **E5**= Performance of Naïve Bayes when, Without Noise + Percentage Split of 20%:80% |
| **E6**= Performance of Naïve Bayes when, Without Noise + Percentage Split of 80%:20% |
| **E7**= Performance of Naïve Bayes when, With Noise + Percentage Split of 20%:80% |
| **E8**= Performance of Naïve Bayes when, With Noise + Percentage Split of 80%:20% |
| **E9**= Performance of JRip when, Without Noise + Percentage Split of 20%:80% |
| **E10**= Performance of Stacking when, Without Noise + Percentage Split of 80%:20% |
| **E11**= Performance of JRip when, With Noise + Percentage Split of 20%:80% |
| **E12**= Performance of JRip when, With Noise + Percentage Split of 80%:20% |

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

As shown by the table above, there are three total classifiers used in this analysis. These classifiers are J48, Naïve Bayes, and JRip. Each classifier is tested with either all 10 or selected attributes. In addition to the Noise, each training and testing set are split either 80%/20% or 20%/80% respectively. Each combination of tests is run 10 times with 10 unique seed values. J48 from E1 – E4, Naïve Bayes from E4 – E8, and JRip from E9 – E12. The highlighted portions of the tables below show us the average accuracy and variance of each tests run. The results of the test run are shown by the tables E1 – E12 below:

|  |  |  |
| --- | --- | --- |
| Table for E2 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 52.8 |
| 2 | 2 | 51.5 |
| 3 | 3 | 50.1 |
| 4 | 4 | 55.4 |
| 5 | 5 | 51 |
| 6 | 6 | 51.7 |
| 7 | 7 | 51.5 |
| 8 | 8 | 51.5 |
| 9 | 9 | 49.5 |
| 10 | 10 | 51.6 |
| Average | 51.66 | |
| Variance | 2.544889 | |

|  |  |  |
| --- | --- | --- |
| Table for E1 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 50.5374 |
| 2 | 2 | 51.2372 |
| 3 | 3 | 49.7626 |
| 4 | 4 | 51.912 |
| 5 | 5 | 49.8875 |
| 6 | 6 | 50.4874 |
| 7 | 7 | 50.9623 |
| 8 | 8 | 51.4871 |
| 9 | 9 | 51.4121 |
| 10 | 10 | 51.2872 |
| Average | 50.89728 | |
| Variance | 0.502118 | |

|  |  |  |
| --- | --- | --- |
| Table for E4 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 50.2 |
| 2 | 2 | 47.6 |
| 3 | 3 | 47.5 |
| 4 | 4 | 52.7 |
| 5 | 5 | 49.6 |
| 6 | 6 | 49.7 |
| 7 | 7 | 48.2 |
| 8 | 8 | 48.1 |
| 9 | 9 | 48.5 |
| 10 | 10 | 48.4 |
| Average | 49.05 | |
| Variance | 2.469444 | |

|  |  |  |
| --- | --- | --- |
| Table for E3 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 48.5379 |
| 2 | 2 | 48.5129 |
| 3 | 3 | 46.4384 |
| 4 | 4 | 48.7628 |
| 5 | 5 | 47.0882 |
| 6 | 6 | 48.063 |
| 7 | 7 | 47.7881 |
| 8 | 8 | 48.2379 |
| 9 | 9 | 48.7878 |
| 10 | 10 | 48.8628 |
| Average | 48.10798 | |
| Variance | 0.639796 | |

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

|  |  |  |
| --- | --- | --- |
| Table for E6 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 54.3 |
| 2 | 2 | 53.9 |
| 3 | 3 | 54.6 |
| 4 | 4 | 54.8 |
| 5 | 5 | 53 |
| 6 | 6 | 52.9 |
| 7 | 7 | 54.1 |
| 8 | 8 | 54 |
| 9 | 9 | 52.8 |
| 10 | 10 | 55.8 |
| Average | 54.02 | |
| Variance | 0.888444 | |

|  |  |  |
| --- | --- | --- |
| Table for E5 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 52.3369 |
| 2 | 2 | 51.5621 |
| 3 | 3 | 53.5116 |
| 4 | 4 | 50.4874 |
| 5 | 5 | 51.6371 |
| 6 | 6 | 51.6121 |
| 7 | 7 | 51.1122 |
| 8 | 8 | 47.813 |
| 9 | 9 | 52.3369 |
| 10 | 10 | 50.562 |
| Average | 51.29713 | |
| Variance | 2.302284 | |

|  |  |  |
| --- | --- | --- |
| Table for E7 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 47.913 |
| 2 | 2 | 45.9385 |
| 3 | 3 | 46.9383 |
| 4 | 4 | 48.188 |
| 5 | 5 | 47.838 |
| 6 | 6 | 49.0377 |
| 7 | 7 | 47.3132 |
| 8 | 8 | 46.2384 |
| 9 | 9 | 48.5629 |
| 10 | 10 | 46.9133 |
| Average | 47.48813 | |
| Variance | 0.997285 | |

|  |  |  |
| --- | --- | --- |
| Table for E8 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 50 |
| 2 | 2 | 50.2 |
| 3 | 3 | 50 |
| 4 | 4 | 50 |
| 5 | 5 | 49.9 |
| 6 | 6 | 50.2 |
| 7 | 7 | 50.3 |
| 8 | 8 | 50.4 |
| 9 | 9 | 49.2 |
| 10 | 10 | 50.3 |
| Average | 50.05 | |
| Variance | 0.116111 | |

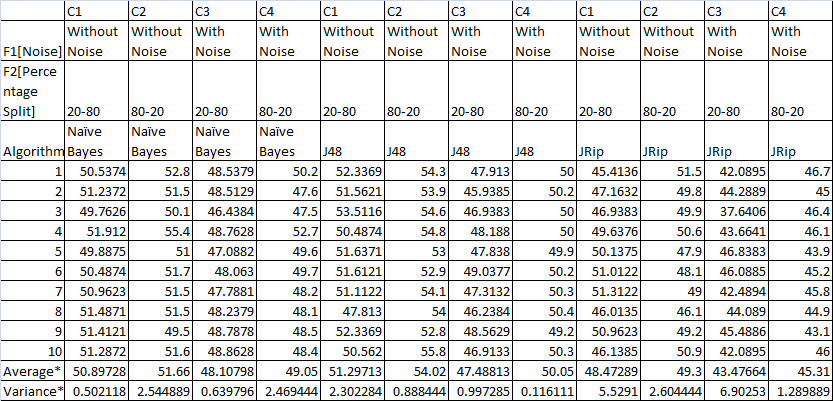
|  |  |  |
| --- | --- | --- |
| Table for E9 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 45.4136 |
| 2 | 2 | 47.1632 |
| 3 | 3 | 46.9383 |
| 4 | 4 | 49.6376 |
| 5 | 5 | 50.1375 |
| 6 | 6 | 51.0122 |
| 7 | 7 | 51.3122 |
| 8 | 8 | 46.0135 |
| 9 | 9 | 50.9623 |
| 10 | 10 | 46.1385 |
| Average | 48.47289 | |
| Variance | 5.5291 | |

|  |  |  |
| --- | --- | --- |
| Table for E10 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 51.5 |
| 2 | 2 | 49.8 |
| 3 | 3 | 49.9 |
| 4 | 4 | 50.6 |
| 5 | 5 | 47.9 |
| 6 | 6 | 48.1 |
| 7 | 7 | 49 |
| 8 | 8 | 46.1 |
| 9 | 9 | 49.2 |
| 10 | 10 | 50.9 |
| Average | 49.3 | |
| Variance | 2.0604444 | |

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

|  |  |  |
| --- | --- | --- |
| Table for E12 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 46.7 |
| 2 | 2 | 45 |
| 3 | 3 | 46.4 |
| 4 | 4 | 46.1 |
| 5 | 5 | 43.9 |
| 6 | 6 | 45.2 |
| 7 | 7 | 45.8 |
| 8 | 8 | 44.9 |
| 9 | 9 | 43.1 |
| 10 | 10 | 46 |
| Average | 45.31 | |
| Variance | 1.289889 | |

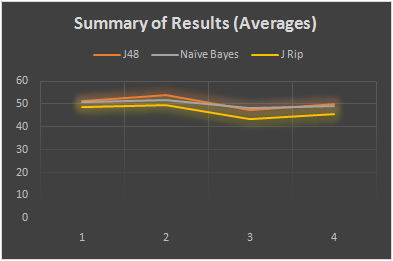
|  |  |  |
| --- | --- | --- |
| Table for E11 | | |
| Trial | Seed | Accuracy |
| 1 | 1 | 42.0895 |
| 2 | 2 | 44.2889 |
| 3 | 3 | 37.6406 |
| 4 | 4 | 43.6641 |
| 5 | 5 | 46.8383 |
| 6 | 6 | 46.0885 |
| 7 | 7 | 42.4894 |
| 8 | 8 | 44.089 |
| 9 | 9 | 45.4886 |
| 10 | 10 | 42.0895 |
| Average | 43.47664 | |
| Variance | 6.90253 | |

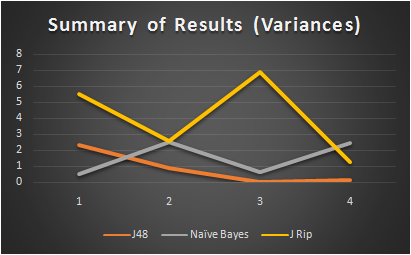


INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**5.1 SUMMARY OF RESULTS**

Here, as expected, C3 and C4 perform relatively poorly.  
This is due to the introduction of noise in the data (class variable).J48 provides the best performance among all the classifiers used to train data with an accuracy of maximum 54.02%. Naive Bayes comes second with a similar but slightly lower accuracy. Our third classifier JRip performs poorly compared to other algorithms.





As shown in the graph above, J48 has the lowest variance from C1 to C4.

The JRip has the highest variation among all classifiers from C1 to C4 ; That is, the JRip is the least stable among all three classifiers.

The JRip classifier varied 5 times between C3 and C4 i.e.: in the presence of noise.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**5. ANALYSIS AND CONCLUSION**

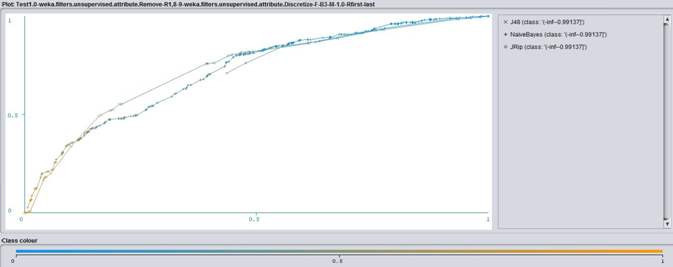
**5.1 ROC CURVE**

**Definition**: A receiver operating characteristic (ROC) is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR

= true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings.

As the definition states, a ROC curve plots the accuracy of a classifier to predict the TPR (True Positive Rate) and FPR (False Positive Rate) on a curve. This results in finding out the accuracy with which our classifiers are able to predict the true positives and true negatives. This method of determining the classifier and factor overall efficiency is by ‘how much area is covered under the ROC curve. Higher the area, better the model.

**Generating Multiple ROC Curves**:We designed a model using the ‘Knowledge Flow’ feature in Weka to draw multiple ROC curves for one factor comparison with another. The figure below shows the flow and design of the experiment replication using the knowledge flow feature:



ROC Curve for **Selected Attributes** (both 20%/80% and 80%/20% split)

A quick analysis of these ROC curves helps us infer that the efficiency of ‘**All Attributes**’ is higher than selected attributes because the *Area under the ROC curve is larger*for all attributes.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**5.2 CLASSIFIER ANALYSIS**

The highest accuracies and lowest variances for all classifiers are stated below:

|  |  |  |
| --- | --- | --- |
| Classifier Name | Highest Accuracy | Lowest Variance |
| Naïve Bayes | 51.66 | 0.502118 |
| J48 | 54.02 | 0.116111 |
| Jrip | 49.3 | 1.289889 |

This table makes it clear that *J48 Decision Tree classifier builds the best model*as it has the best accuracy and lowest variance (best stability) among all the factor and criteria combination.

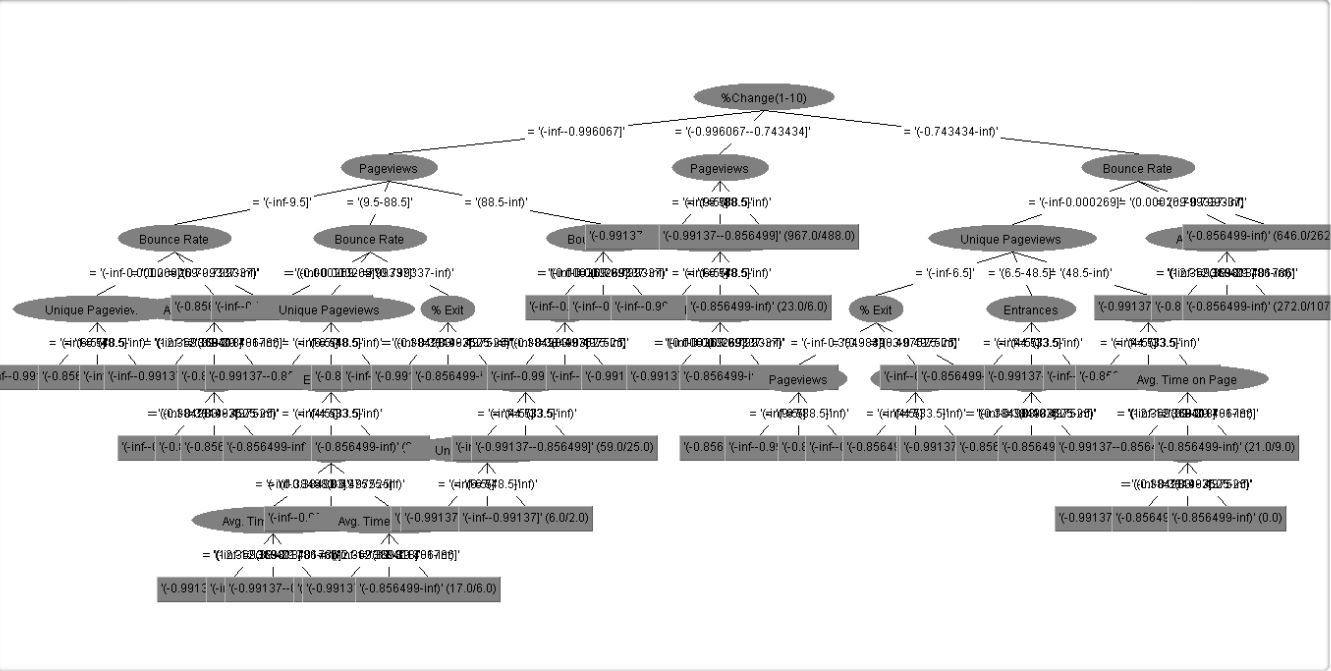
=== Confusion Matrix === J48

a b c <-- classified as

309 80 162 | a = '(-inf--0.99137]'

212 164 190 | b = '(-0.99137--0.856499]'

112 110 361 | c = '(-0.856499-inf)'



INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

**5.3 ATTRIBUTE ANALYSIS**

We used two sets of attributes as a factor (Number of attributes) in our experiment. We need to be able to infer which of these two attribute sets produce a better model for predicting the attribute class. We conclude that*‘All attributes’ set builds a better model*because of the two reasons stated below:

1. The ROC curve analysis in section 5.1 has concluded that the All Attributes set with various combinations of factors produces a better area under the ROC curve i.e. it is better able to predict the TP (Tue Positives) and TN (True Negatives), in turn keeping the FP and FN count down.
2. Observing the results for average accuracy and variance we observe that the accuracy increases by about 4% when the All Attributes set is being used. This also suggests that the algorithms are able to build a better predictive model when supplied with all attributes in the experiment.

**5.4 CONCLUSION**

With the average accuracy and variance, ROC curves, Attribute and Classifier evaluation we are recommending the following for our Wholesaler dataset:

1. **Discretization**: Discretization of *Annual Premium*attribute improved the model overall accuracy across classifiers.
2. **Classifier**: *Naïve Bayes*has performed with highest accuracy and most stability with this dataset.
3. **Number of Attributes factor**: Although we employed two methods (*OneR*and C*lassifierSubsetEval*) to determine ‘Selected Attributes’, a separate combination among all the attributes emerged to be more accurate.
4. **Percentage split factor:** An *80/20 percentage split*in training/test data comes up with a better prediction model than vice versa.

INSY 5339 ***Principles of Business Data Mining*** Dr. Sikora

# **6. REFRENCES**

* Ian H. Witten; Eibe Frank; Mark A. Hall (2011). ["Data Mining: Practical machine learning tools and techniques, 3rd Edition"](http://www.cs.waikato.ac.nz/~ml/weka/book.html). Morgan Kaufmann, San Francisco. Retrieved 2011-01-19.
* WAA Standards Committee. "Web analytics definitions." Washington DC: Web Analytics Association (2008)
* H. Jantan et al.“Classification for Prediction”,International Journal on Computer Science
* and Engineering, 2(8): 2526-2534, 2010
* **Data Cleaning** [Online] / auth. Authors Multiple // Wikipedia. - 5 1, 2013. - [http://en.wikipedia.org/wiki/Data\_cleansing.](http://en.wikipedia.org/wiki/Data_cleansing)
* **Google Refine** [Online] / auth. Authors Multiple // Google. - Google Refine, 7 1, 2013. - https://code.google.com/p/google-refine/.
* **Making Sense of Data** [Online] / auth. Authors Multiple // Google . - Google, 3 1, 2014. - https://datasense.withgoogle.com/course.
* **Multiple ROC Curves** [Online] / auth. Shams Rushdi // Youtube. - 10 28, 2013. - https:[//www.youtube.com/watch?v=rZHw3gGe7DA.](http://www.youtube.com/watch?v=rZHw3gGe7DA)