INSY 5339 – Principles of Business Data Mining

Project Report:
San Francisco Crime Classification

Group 2:
Arbaaz Shaikh
Gulrej Muradali Hamirani
Padmavathi Karunaiananda Sekar
Sanchit Puri
Venkatesh Singanallur Ramanathan

TABLE OF CONTENTS

1. DATA BACKGROUND

Background

During the period from 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inevitable island of Alcatraz.

Unlike today, the city was previously known more for its crime activity than technology it, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no paucity of crime in the city by the bay.

From Sunset to SOMA, and Marina to Excelsior, this dataset provides a total of 12 years of crime reports from across all of San Francisco's neighborhoods

Objective

To predict the category of crimes that occurred in the city by the bay given the time and location.

Dataset & Attributes

This dataset is taken from Kaggle.

Source: https://www.kaggle.com/c/sf-crime/data

Kaggle (https://www.kaggle.com/about) is a platform dedicated to data mining. Particularly, it hosts public data science challenges, the problems/challenges are posted by sponsors, and people across the globe can compete for putting together the best solution.

This dataset is brought to you by **SF OpenData**, the central clearinghouse for data published by the City and County of San Francisco. It contains crime reports from January 2003 to May 2015.

Size of the data: 878,049 Records (12 years of data)

Attributes Description:

The dataset that was provided had 9 attributes and 878, 049 instances.

The attributes are as below:

Attribute Name	Description
Dates	Timestamp of the crime incident.
Descript	Detailed description of the crime incident.
DayOfWeek	The day of the week of the crime incident.
PdDisrict	Name of the Police Department District for the incident.
Resolution	How the crime incident was resolved.

Address	The approximate street address of the crime incident.
X	Longitude
Y	Latitude
Category	Category of the crime incident. This is the **target variable** you are going to predict.

1.3 Class Attribute

In the dataset, there were total of 39 categories of crime which were to be predicted based on attribute time and location. These categories of crime are nominal values.

Below is the list of 39 categories of crimes:

Larceny/theft
Other Offenses
Non criminal
Assault
Posession of Drug/Narcotic
Vehicle Theft
Vandalism
Warrants
Burgulary
Suspicious Occ
Missing person
Robbery
Fraud
Forgery/Counterfiet
secondary codeds
weapons law
Prostitution
Tresspassing
Stolen Property
Sex offences(forcible)
Disorderly Conduct
Drunkenness(influence in public place)
Recorded Vehicle
Kidnapping
Driving under the influence
Runaway
Liquor laws
Arson
loitering
Embezzlement(Grand Theft by an employee)

Suicide
Family Offenses
Bad Checks
Bribery
Extortion
Sex offenses
Gambling
Pornography/Obsene
Treaspassing

2. Data Cleaning Process

It is a process of preparation of clean data set given inconsistent, erroneous and missing data. It involves identifying problematic records and correcting them to improve Data Quality. Data Cleansing is a crucial process in Data Mining which is required to gain good accuracy. There are several ways to perform Data Cleansing. The approach followed in our project are mentioned below:

- Retrieve data from source (Kaggle website)
- Identify problematic records – missing data, inconsistent data, and irrelevant data
- Remove or modify problematic records
- Run algorithms in Weka to check if the clean dataset is improving prediction accuracy

The tool we used for Data Cleansing is **Microsoft Excel**. This tool gave us the flexibility needed for handling our huge dataset.

The original list of attributes in our dataset are:

Name of the Attribute	Description of the Attribute
Dates	Timestamp of the crime incident
Category	Category of the crime incident. This is the target variable we are going to predict
Descript	Detailed description of the crime incident
DayOfWeek	The day of the week of the crime incident
PdDistrict	Name of the Police Department District for the incident
Resolution	How the crime incident was resolved
Address	The approximate street address of the crime incident
X	Longitude
Y	Latitude

	A	B	C	D	E	F	G	H	I
1	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
2	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258916751	37.7745985957
3	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258916751	37.7745985957
4	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.4243630215	37.800414322
5	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.4269953268	37.8008726328
6	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.4387376228	37.7715411721
7	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM UNLOCKE	Wednesday	INGLESIDE	NONE	0 Block of TEDDY AV	-122.4032523612	37.7134307041
8	2015-05-13 23:30:00	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	INGLESIDE	NONE	AVALON AV / PERU AV	-122.4233269767	37.7251380404
9	2015-05-13 23:30:00	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	BAYVIEW	NONE	KIRKWOOD AV / DONAHUE ST	-122.3712743174	37.727564072
10	2015-05-13 23:00:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	RICHMOND	NONE	600 Block of 47TH AV	-122.5081940311	37.7766012607
11	2015-05-13 23:00:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAVENWORTH ST	-122.4190876767	37.8078015517
12	2015-05-13 22:58:00	LARCENY/THEFT	PETTY THEFT FROM LOCKED AU	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAVENWORTH ST	-122.4190876767	37.8078015517
13	2015-05-13 22:30:00	OTHER OFFENSES	MISCELLANEOUS INVESTIGATIO	Wednesday	TARAVAL	NONE	0 Block of ESCOLTA WY	-122.4879830728	37.7376666543
14	2015-05-13 22:30:00	VANDALISM	MALICIOUS MISCHIEF, VANDALI	Wednesday	TENDERLOIN	NONE	TURK ST / JONES ST	-122.4124142636	37.7830037965
15	2015-05-13 22:06:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	NORTHERN	NONE	FILLMORE ST / GEARY BL	-122.4329146035	37.7843533427
16	2015-05-13 22:00:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	BAYVIEW	NONE	200 Block of WILLIAMS AV	-122.3977444271	37.7299346936
17	2015-05-13 22:00:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	BAYVIEW	NONE	0 Block of MENDELL ST	-122.383691504	37.743189042
18	2015-05-13 22:00:00	ROBBERY	ROBBERY, ARMED WITH A KNIFI	Wednesday	TENDERLOIN	NONE	EDDY ST / JONES ST	-122.4125973772	37.7839320277
19	2015-05-13 21:55:00	ASSAULT	AGGRAVATED ASSAULT WITH B	Wednesday	INGLESIDE	NONE	GODEUS ST / MISSION ST	-122.4216815316	37.7428222005
20	2015-05-13 21:40:00	OTHER OFFENSES	TRAFFIC VIOLATION	Wednesday	BAYVIEW	ARREST, BOOKED	MENDELL ST / HUDSON AV	-122.38640087	37.7389834911
21	2015-05-13 21:30:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	TENDERLOIN	NONE	100 Block of JONES ST	-122.4122497676	37.7825563302
22	2015-05-13 21:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	INGLESIDE	NONE	200 Block of EVELYN WY	-122.4493891113	37.7426688026
23	2015-05-13 21:17:00	ROBBERY	ROBBERY, BODILY FORCE	Wednesday	INGLESIDE	NONE	1600 Block of VALENCIA ST	-122.4202721353	37.7473316299
24	2015-05-13 21:11:00	WARRANTS	WARRANT ARREST	Wednesday	TENDERLOIN	NONE	100 Block of JONES ST	-122.4122497676	37.7825563302
25	2015-05-13 21:11:00	NON-CRIMINAL	STAY AWAY OR COURT ORDER, P	Wednesday	TENDERLOIN	NONE	100 Block of JONES ST	-122.4122497676	37.7825563302
26	2015-05-13 21:10:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AI	Wednesday	NORTHERN	NONE	FILLMORE ST / LOMBARD ST	-122.4360492036	37.7998412229
27	2015-05-13 21:00:00	NON-CRIMINAL	LOST PROPERTY	Wednesday	TENDERLOIN	NONE	300 Block of OFARRELL ST	-122.4105092588	37.7860432223

The data cleaning done on our dataset involves removal of irrelevant attributes and false predictors and deriving new attributes.

2.1 Irrelevant Attributes

These are the attributes that do not contribute to the prediction of the class attribute. Retaining the irrelevant attributes in the dataset will lead to lower prediction accuracy. Some of the irrelevant attributes in our dataset are:

I. **Resolution**: This attribute gives information about the action taken on the crime reported. In the original dataset, we see a lot of values with null for this attribute. Since this attribute does not contain information that can be used to predict the class variable, we remove this attribute.

II. **Address**: This attribute gives information about the street address where the crime took place. Although this attribute can be used to predict the class attribute, without the postal code in the address, this attribute is not of use in our dataset.

III. **X – Longitude**: This attribute gives information about the location (Longitude) where the crime happened. This attribute can be used to derive the postal code attribute and with this new attribute it is possible to predict the class variable. Deriving this new attribute is out of scope due to the complexity involved in retrieving the required information from the geo-tracker.

IV. **Y – Latitude**: This attribute gives information about the location (Latitude) where the crime happened. This attribute can be used to derive the postal code attribute and with this new attribute it is possible to predict the class variable. Deriving this new attribute is out of scope due to the complexity involved in retrieving the required information from the geo-tracker.

2.2 False Predictors

These are attributes which predict the class variable too well such that the accuracy shoots up. False predictors in the dataset should be removed to check how well the algorithm performs in predicting the class variable. In our dataset, we have one false predictor attribute:

I. **Descript**: This attribute provides detail information about the crime that occurred. When this variable is used to predict the algorithm, we see that certain words from this attribute are used to predict the class variable. The accuracy is as much as 99.9% when the dataset with this variable is used. Hence, this attribute should be removed for accurate prediction of the class attribute.

2.3 Deriving New Attributes

These are attributes that are derived from the existing attributes in the dataset which are found to help in predicting the class attribute. In our project, we have derived a few attributes as mentioned below:

I. **Dates** – This attribute is split to obtain attributes **Time** and **Date**.
II. **Time** – This derived attribute is further used to derive **TimeSplit**. The "TimeSplit" attribute defines the time interval in the day when the crime occurred. This attribute has values –
    - *Morning*
    - *Afternoon*
    - *Evening*
    - *Night*
III. **Date** – This derived attribute is further used to derive **Month**, **Year** and **DayOfMonth**. The "Month" contains information about the month when the crime occurred. It takes 12 values which are the months of the year. The "Year" attribute contains information about the year when the crime occurred. It takes 12 values which are the years when data was collected (2003-2015). The "DayOfMonth" attribute contains information about the day in the month when the crime was reported. It contains 31 values which are the days of the month.

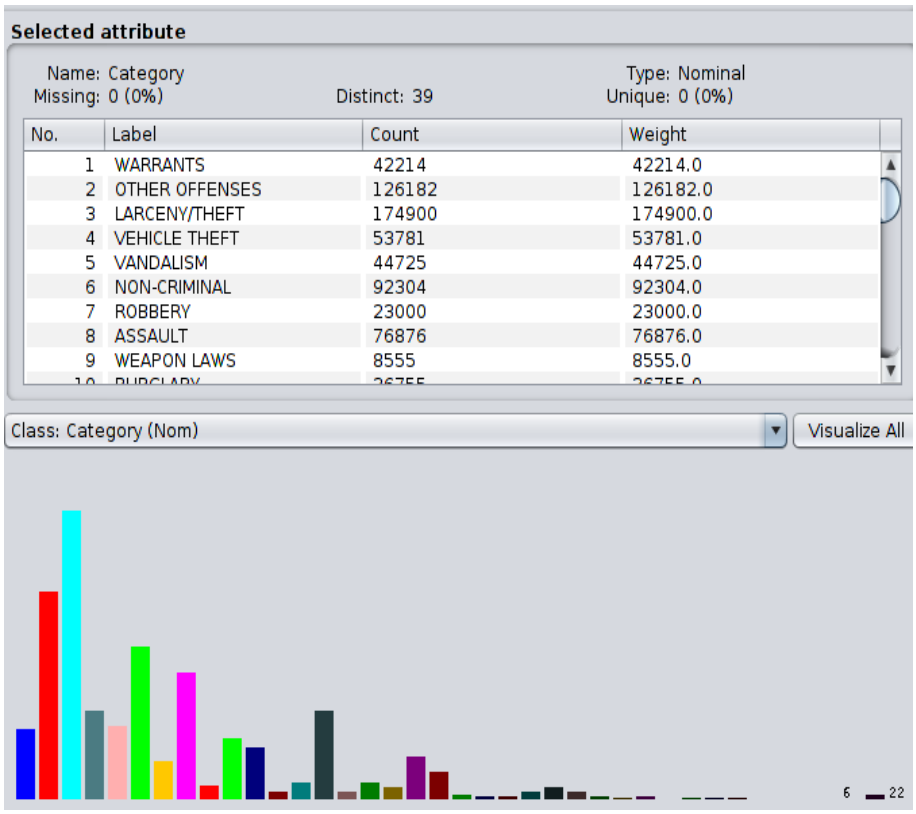
After cleansing, we have the following list of attributes in our dataset:

- Year
- DayOfMonth
- Month
- TimeSplit
- DayOfWeek
- PdDistrict
- Category

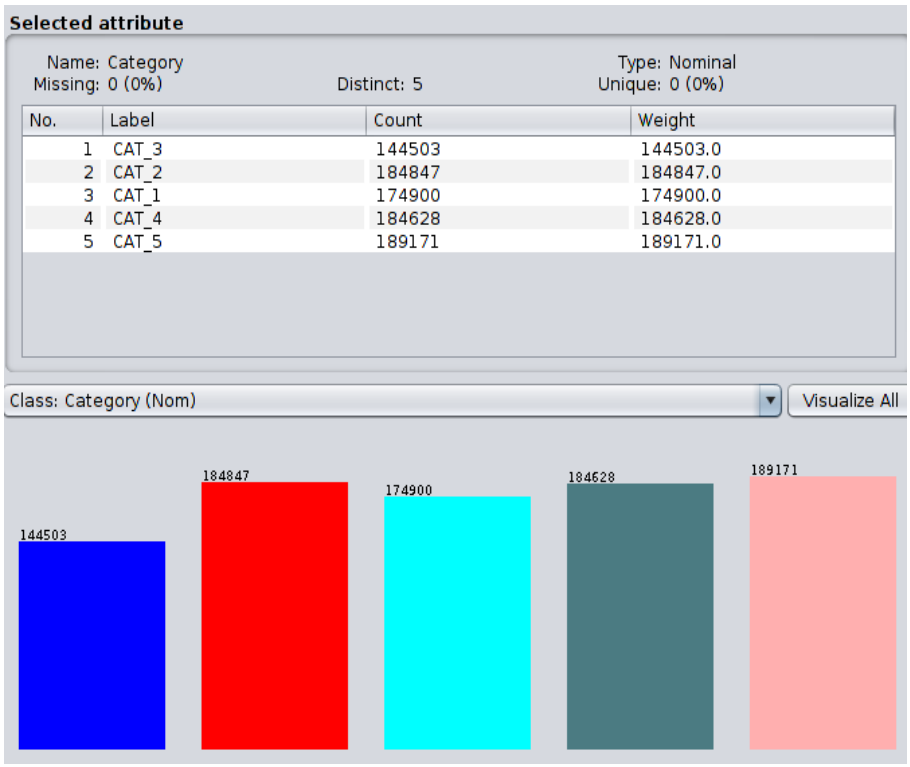
A	B	C	D	E	F	G
Year	**DayOfMonth**	**Month**	**Time Split**	**DayOfWeek**	**PdDistrict**	**Category**
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_3
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_2
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_2
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_1
2015	13	5	EVENING	Wednesday	PARK	CAT_1
2015	13	5	EVENING	Wednesday	INGLESIDE	CAT_1
2015	13	5	EVENING	Wednesday	INGLESIDE	CAT_4
2015	13	5	EVENING	Wednesday	BAYVIEW	CAT_4
2015	13	5	EVENING	Wednesday	RICHMOND	CAT_1
2015	13	5	EVENING	Wednesday	CENTRAL	CAT_1
2015	13	5	EVENING	Wednesday	CENTRAL	CAT_1
2015	13	5	EVENING	Wednesday	TARAVAL	CAT_2
2015	13	5	EVENING	Wednesday	TENDERLOIN	CAT_5
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_1
2015	13	5	EVENING	Wednesday	BAYVIEW	CAT_3
2015	13	5	EVENING	Wednesday	BAYVIEW	CAT_3
2015	13	5	EVENING	Wednesday	TENDERLOIN	CAT_5
2015	13	5	EVENING	Wednesday	INGLESIDE	CAT_4
2015	13	5	EVENING	Wednesday	BAYVIEW	CAT_2
2015	13	5	EVENING	Wednesday	TENDERLOIN	CAT_3
2015	13	5	EVENING	Wednesday	INGLESIDE	CAT_1
2015	13	5	EVENING	Wednesday	INGLESIDE	CAT_5
2015	13	5	EVENING	Wednesday	TENDERLOIN	CAT_3
2015	13	5	EVENING	Wednesday	TENDERLOIN	CAT_3
2015	13	5	EVENING	Wednesday	NORTHERN	CAT_1

2.4 Balancing the Dataset

As part of data cleansing, we also had to handle skewed data and make changes to balance out the skewed dataset. To do this, we condensed the 39 values of the category attribute into 5 categories. We also condensed the 8 values of the TimeSplit attribute into 4 values. The following graphs depict the distribution of the dataset for the **Category** attribute:



Original Dataset showing 39 Categories



Balanced Dataset showing 5 Categories

3. EXPERIMENT DESIGN

CLASSIFIER SELECTION

We selected the following classifiers for our experiment design:

1. **J48 (tree)** - J48, a tree classifier optimizes the most efficient attribute which increases the prediction accuracy. The normalized information gain is the criteria for splitting of nodes. The highest normalized information gain is chosen to be decision attribute.
2. **Naïve Bayes (Bayes)** – The Naïve Bayes classifier is based on the Bayes rule of conditional probability to determine the attributes upon which model is to be built. This approach is often able to produce a highly stable model because the nature of data does not change drastically in most cases.
3. **Decision Table** –Decision tables are used to model complicated logic. It is used to detect combinations of conditions. A decision table comprises of a hierarchical table in which higher level entry table gets broken down by the values of additional attributes to form another table.
4. **Zero R** - ZeroR is the simplest classification method which relies on the target and ignores all other attributes. It is useful for determining a benchmark for other classification methods. (21.059% in our case).

FOUR CELL EXPERIMENT DESIGN

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

1. **Factor 1 (F1) → No Noise vs 10% Noise**
2. **Factor 2 (F2) → Percentage Split**

Four 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 more clearly in the table blow.

	No Noise	10% Noise
Original Dataset	C1	C3
Aggregated Dataset	C2	C4

1. F11, C1= Original Dataset without Noise
2. F12, C2= Original Dataset with Noise
3. F21, C3= Aggregated Dataset without Noise
4. F22, C2= Aggregated Dataset with Noise

4. EXPERIMENT RESULTS

Results for Each Classifier:

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 Naïve Bayes when, Original Dataset without Noise + Percentage Split of 50%:50%
E2= Performance of Naïve Bayes when, Original Dataset with Noise + Percentage Split of 50%:50%
E3= Performance of Naïve Bayes when, Aggregated Dataset without Noise + Percentage Split of 50%:50%
E4= Performance of Naïve Bayes when, Aggregated Dataset with Noise + Percentage Split of 50%:50%
E5= Performance of Decision Table when, Original Dataset without Noise + Percentage Split of 50%:50%
E6= Performance of Decision Table when, Original Dataset with Noise + Percentage Split of 50%:50%
E7= Performance of Decision Table when, Aggregated Dataset without Noise + Percentage Split of 50%:50%
E8= Performance of Decision Table when, Aggregated Dataset with Noise + Percentage Split of 50%:50%
E9= Performance of J48 when, Original Dataset without Noise + Percentage Split of 50%:50%
E10= Performance of J48 when, Original Dataset with Noise + Percentage Split of 50%:50%
E11= Performance of J48 when, Aggregated Dataset without Noise + Percentage Split of 50%:50%
E12= Performance of J48 when, Aggregated Dataset with Noise + Percentage Split of 50%:50%

As shown in the table above, there are three total classifiers used in this analysis. These classifiers are Naïve Bayes, Decision Table and J48. Each classifier is tested on the original dataset (39 attributes) with and without noise and on the aggregated dataset (5 attributes) with and without noise. In addition to the attribute selection, each training and testing set are split 50%/50% respectively. Each combination of tests is run 10 times with 10 unique seed values. Naïve Bayes from E1 – E4, Decision Table from E5 – E8, and J48 from E9 – E12. The highlighted portions of the table below show us average accuracy and variance on each test run.

The results of the test run are shown by the tables E1 – E12 below:

Table for E1

Trail	Seed	Accuracy
1	1	22.5891
2	2	22.5550
3	3	22.5764
4	4	22.5981
5	5	22.6154
6	6	22.5477
7	7	22.6434
8	8	22.5382
9	9	22.6199
10	10	22.5099
Average		**22.5793**
Variance		**0.0416**

Table for E2

Trail	Seed	Accuracy
1	1	20.5303
2	2	20.3907
3	3	20.5561
4	4	20.5565
5	5	20.5766
6	6	20.4964
7	7	20.6002
8	8	20.5048
9	9	20.5731
10	10	20.4784
Average		**20.5263**
Variance		**0.0614**

Table for E3

Trail	Seed	Accuracy
1	1	28.5082
2	2	28.4683
3	3	28.4925
4	4	28.4483
5	5	28.5041
6	6	28.4310
7	7	28.4716
8	8	28.4670
9	9	28.4454
10	10	28.4415
Average		**28.4674**
Variance		**0.02690**

Table for E4

Trail	Seed	Accuracy
1	1	27.4256
2	2	27.4320
3	3	27.4051
4	4	27.4074
5	5	27.4318
6	6	27.3523
7	7	27.4668
8	8	27.4395
9	9	27.4058
10	10	27.4274
Average		**27.4194**
Variance		**0.03000**

Table for E5

Trail	Seed	Accuracy
1	1	22.8521
2	2	22.8329
3	3	22.8591
4	4	22.9798
5	5	22.9120
6	6	22.8593
7	7	22.0053
8	8	22.8564
9	9	22.8983
10	10	22.8334

Average **22.8889**
Variance **0.0604**

Table for E6

Trail	Seed	Accuracy
1	1	20.8098
2	2	20.7868
3	3	20.8390
4	4	20.7989
5	5	20.8068
6	6	20.7351
7	7	20.8686
8	8	20.7529
9	9	20.7458
10	10	20.7007

Average **20.7844**
Variance **0.0509**

Table for E7

Trail	Seed	Accuracy
1	1	29.3168
2	2	29.1954
3	3	29.2642
4	4	29.9945
5	5	29.1991
6	6	29.1784
7	7	29.1501
8	8	29.1991
9	9	29.2440
10	10	29.2196

Average **29.1961**
Variance **0.00851**

Table for E8

Trail	Seed	Accuracy
1	1	27.8065
2	2	27.8304
3	3	27.1169
4	4	27.0349
5	5	27.0513
6	6	27.9916
7	7	27.0786
8	8	27.0137
9	9	27.8673
10	10	27.0638

Average **27.9855**
Variance **0.1104**

Table for E9

Trail	Seed	Accuracy
1	1	21.1214
2	2	21.2480
3	3	21.1731
4	4	21.1323
5	5	21.1423
6	6	21.1489
7	7	21.1221
8	8	21.1293
9	9	21.1351
10	10	21.1691
Average		**21.1522**
Variance		**0.03810**

Table for E10

Trail	Seed	Accuracy
1	1	19.0352
2	2	19.0752
3	3	19.0085
4	4	19.0743
5	5	19.0435
6	6	19.0543
7	7	19.0331
8	8	19.0775
9	9	19.0489
10	10	19.0589
Average		**19.0589**
Variance		**0.01921**

Table for E11

Trail	Seed	Accuracy
1	1	27.3851
2	2	27.4402
3	3	27.3413
4	4	27.3477
5	5	27.4577
6	6	27.2131
7	7	27.4453
8	8	27.3432
9	9	27.3756
10	10	27.3468
Average		**27.3889**
Variance		**0.04960**

Table for E12

Trail	Seed	Accuracy
1	1	25.9867
2	2	26.0455
3	3	25.9779
4	4	26.0121
5	5	25.9932
6	6	26.0111
7	7	26.0213
8	8	25.9912
9	9	26.0041
10	10	26.0283
Average		**26.0071**
Variance		**0.02080**

Summary of Results:

The average of the accuracies tested for the three classifiers are shown in the table below:

Algorithm	C1	C2	C3	C4
Naïve Bayes	**22.5793**	**20.5263**	**28.4674**	**27.4194**
Decision Table	**22.8889**	**20.7884**	**29.1961**	**27.9855**
J48	**21.1522**	**19.0589**	**27.3889**	**26.0071**

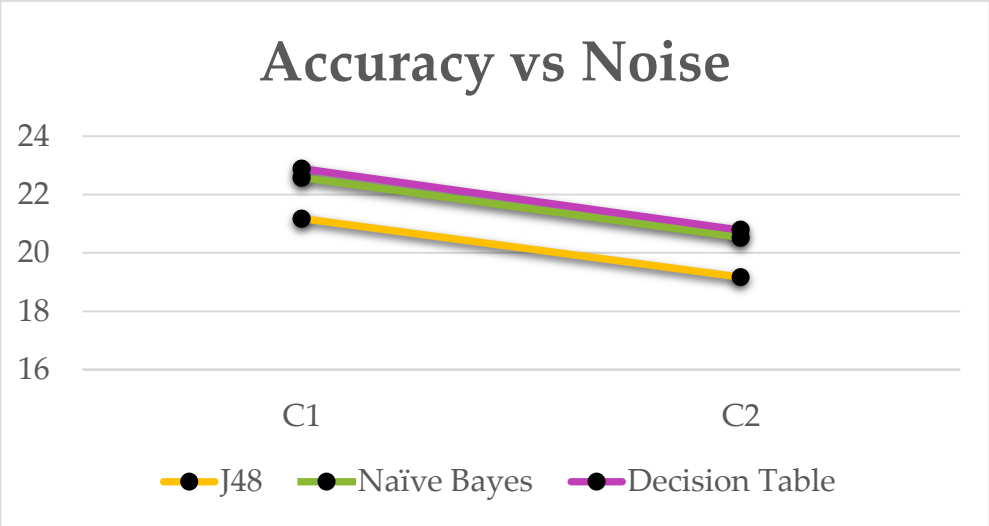


Figure 1: Effect of Noise on Original Dataset

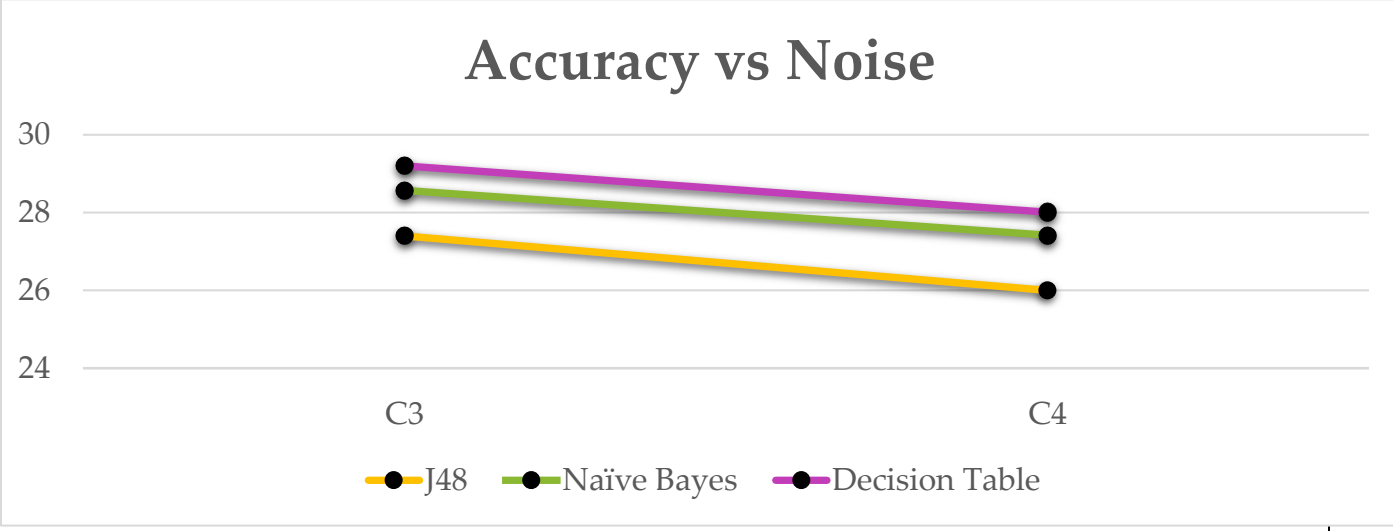


Figure 2: Effect of Noise on Aggregated Dataset

From the above accuracy results, the algorithms for the aggregated dataset with and without noise (C3 & C4) are doing better than original dataset with and without noise (C1 & C2). This was expected because C1 and C2 are having imbalance in their attributes which makes the prediction for the classifiers harder without the imbalance being resolved. The decision table algorithm (29.1961 & 27.9855) has the best accuracy compared to the other algorithms. This is because the algorithm classifies the attributes using a rule based approach. J48, on the other hand, has the worst accuracy because some of the attributes take multiple values and J48 must create individual branches for each attribute, thereby making the classifier complex

to predict and bringing the accuracy down. Naïve Bayes uses a probabilistic approach in predicting the attributes.

The variance tested for the three classifiers are shown in the table below:

Algorithm	C1	C2	C3	C4
Naïve Bayes	0.0416	0.0614	0.02690	0.03000
Decision Table	0.0604	0.0509	0.00851	0.1104
J48	0.03810	0.01921	0.04960	0.02080

From the above table, the aggregated dataset without noise (C3) has the lowest variance (0.00851). This tells us that the decision table algorithm is the most reliable algorithm in terms of stability. The Naïve Bayes algorithm has the second lowest variance making it the next stable algorithm after decision table. J48 has the highest variance therefore not making reliable in terms of stability.

5. ANALYSIS & CONCLUSION

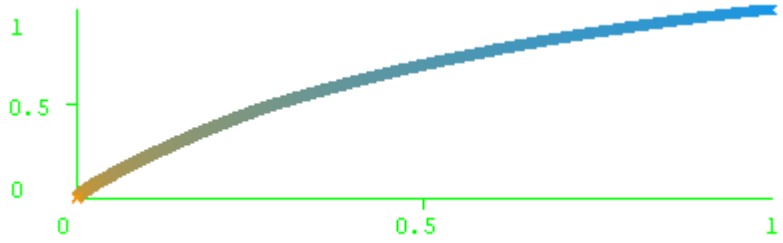
5.1 ROC CURVE

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 can 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.

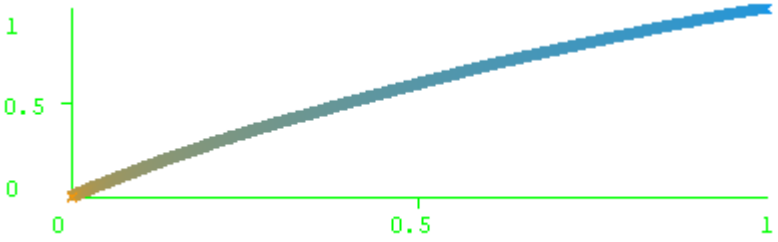
ROC Curve

Dataset: 4 intervals 5 categories
Algorithm: Naïve Bayes
Category - 1



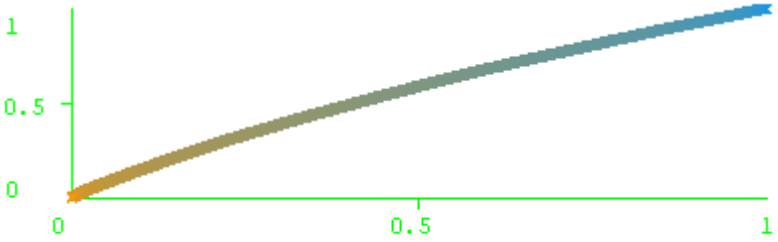
Plot (Area under curve ROC – 0.6433)

Category – 2



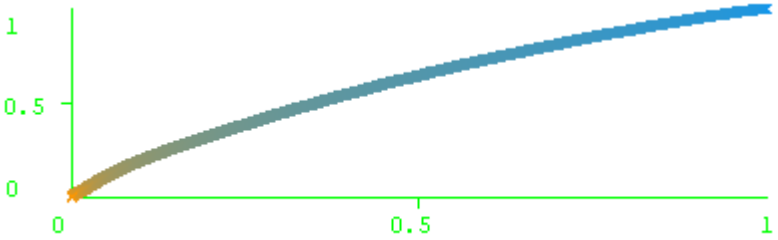
Plot (Area under curve ROC – 0.5697)

Category – 3



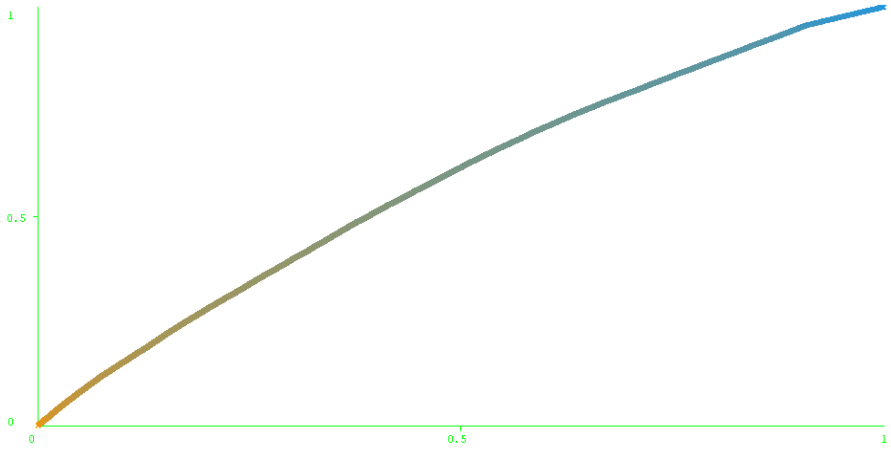
Plot (Area under curve ROC – 0.5598)

Category – 4



Plot (Area under curve ROC – 0.602)

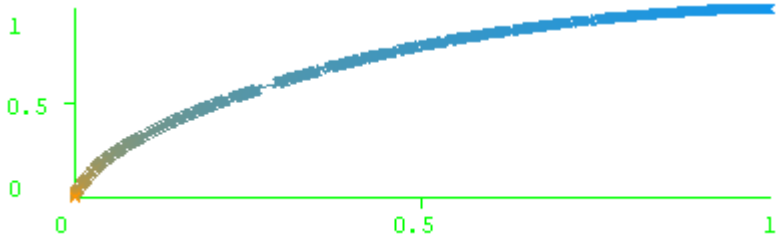
Category – 5



Plot (Area under curve ROC – 0.5788)

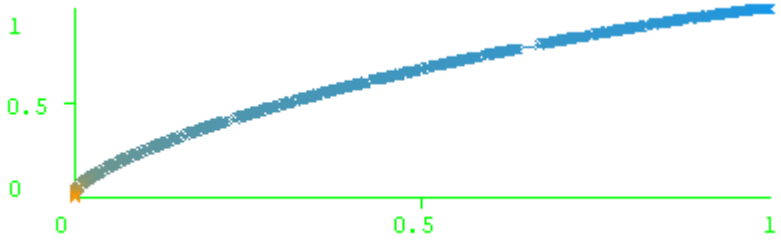
Algorithm: Decision Table

Category – 1



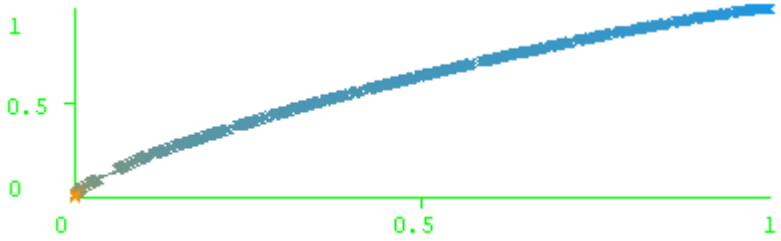
Plot (Area Under Curve ROC -0.7211)

Category – 2



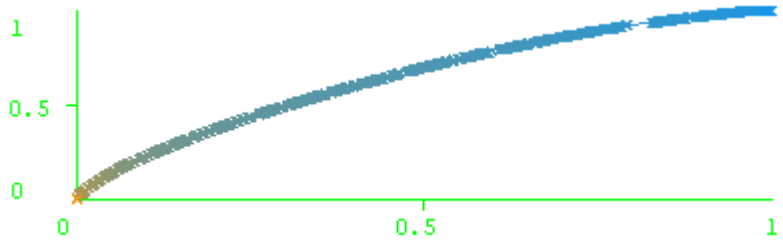
Plot (Area under curve ROC – 0.6349)

Category – 3



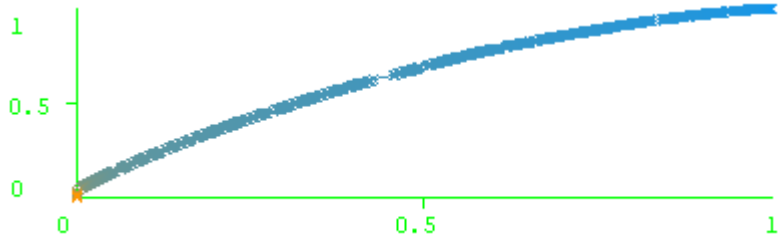
Plot (Area under curve ROC -0.6067)

Category - 4



Plot (Area under curve ROC-0.6422)

Category - 5



Plot (Area under curve ROC-0.6355)

5.2 CLASSIFIER ANALYSIS

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

Classifier Name	Highest Accuracy	Lowest Variance
Decision Table	29.1961	0.00851
J48	27.3889	0.04960
Naïve Bayes	28.4647	0.0614

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

5.3 CONCLUSION

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

- **Data Aggregation**: The Class variable should be aggregated to 5 categories based on relevance then based on number of records.

- **Time Split**: (8 vs 4) A time split of 4 rather than 8 contributed towards getting a better accuracy.
- **Percentage Split:** After testing for a percentage split starting from 10% to 50%, we found out that the overall accuracy across classifiers doesn't improve beyond a 50%split.
- **Classifier:** *Decision Table* has performed with highest accuracy and most likely to correctly predict the category of crime.

6. REFERENCES

[1] Weka Documentation, *http://www.cs.waikato.ac.nz/ml/weka/documentation.html*
[2] San Francisco Crime Classification, *https://www.kaggle.com/c/sf-crime*
[3] Excel Help, *https://support.office.com/en-us/excel*
[4] ROC curves and Area Under the Curve explained (video), *http://www.dataschool.io/roc-curves-and-auc-explained/*