CIS 9557

**Business Analytics**

**Patterns and Trends**

**of**

**Criminal Activity**

**in**

**New York City**

**Group 2**

Madhuri Goenka

Aidel Kagan

Raad Khan

Kishan Patel

Anika Pancho

# **Table of Content**

[**Business Problem** 2](#_Toc58536058)

[**Data Understanding** 2](#_Toc58536059)

[**Data Pre-processing** 3](#_Toc58536060)

[**Regression Analysis** 3](#_Toc58536061)

[**Feature selection/engineering** 4](#_Toc58536062)

[**Model evaluation** 4](#_Toc58536063)

[**Scoring the dataset** 4](#_Toc58536064)

[**Business Insights and Recommendations** 5](#_Toc58536065)

[**Data Limitations and Ethical Issues** 5](#_Toc58536066)

[Appendix 1: Count of Crimes Vs. Classification 7](#_Toc58536067)

[Appendix 2: Pivot of Data for understanding 7](#_Toc58536068)

[Appendix 3: Count of crimes per Borough 8](#_Toc58536069)

[Appendix 4 : Tableau Dashboard 9](#_Toc58536070)

[Appendix 5: Total Crimes vs Temperature 9](#_Toc58536071)

[Appendix 6: Regression Analysis 10](#_Toc58536072)

[Appendix 7: RAPIDMIINER Process flow 10](#_Toc58536073)

[Appendix 8: Attribute Selection 11](#_Toc58536074)

[Appendix 9: Final Model 11](#_Toc58536075)

[APPENDIX 10 :First Confusion Matrix 12](#_Toc58536076)

[APPENDIX 11: Final Confusion Matrix 12](#_Toc58536077)

**Business Problem**

Crime is a social problem affecting many areas of life, including public safety, quality of life, and the economy. By identifying the attributes and parameters associated with criminal activity, we hope to better understand crime patterns and predict trends in criminal activity in New York City. In particular, we are focusing on attributes that may cause differences in level and type of offense and criminal activities occuring during the day, in comparison to night. It’s also imperative to identify the correct type of crime as preventive measures differ for each type of crime. Punishment for different types of crimes is different as well as safeguards need to be put in place to protect the innocent. We want to use this understanding to improve public safety initiatives and enhance efficiency in law enforcement agencies.

# **Data Understanding**

The data shows details of complaints presented to NYPD in 2019. The data includes date, time, location, and type of crime along with suspect and victim age, sex and race attributes. Each reported incident has been broken down according to the level of offense, such as felony, misdemeanor, or violation, along with a granular description of crime corresponding to the internal classification code of the NYPD. The personal attributes help us to identify common parameters within the data and highlight relations of identity in crimes that occurred. We incorporated another dataset of average daily temperature (°F) to identify if weather is a correlating factor in reported crimes.

With the help of Pivot tables and visuals we get the better understanding of data using variations. (Appendix-1,2,3,4) Further investigation can be useful to identify what actions can be taken to reduce the crime and increase the public safety.

# **Data Pre-processing**

We worked on data from the year 2019 as 2020 data had many anomalies given the effects of the COVID-19 pandemic and political unrest. We removed 9 records of data from 6,244, which is roughly 0.15%. Only the data that was meaningless was removed. The changes made included:

* 8 rows were removed since no borough had been specified
* Each day’s average temperature was integrated from a different dataset by indexing date using the vlookup formula
* The combination field for latitude & longitude was removed since the information was redundant (both attributes are separately available).
* 1 row was removed as the age was recorded as 928 years
* Blank cells values shown in the suspect Age/Race/Sex group were changed to “UNKNOWN”
* An additional column was created using the LOOKUP function to identify Day and Night based on the time, with day being 6 AM to 6 PM.

**Regression Analysis**

To further our analysis of the data, we performed regression analysis on it. Regression analysis is a statistical set of processes that assist in determining the relationships between a dependent variable and one or more independent variables.

In our hypothesis, we wanted to test if a change in temperature causes any changes in the amount of crimes that are committed. The first measure we looked at was our adjusted R square value. With only a 3.1% for our R-square value, the regression does not confirm a strong relationship between our variables. The second measure we looked at was our P-value. This tells us that for every single independent variable, do they significantly influence the output or not. Since the p-value was more than 0.05, we know that the temperature does not influence crimes committed.

When we put the data into a scatter plot, it looks more like a bell curve where crimes are committed more during moderate temperatures and not as much for really high or low temperatures. At 41°F we saw a huge jump in the number of crimes committed at a whopping value of 809 crimes. Although this could be seen as an outlier if it was just one value, because it is a sum of all the crimes we found in our dataset at that temperature, we cannot dismiss this finding.

**Feature selection/engineering**

For model building we have utilized the RapidMiner software as a tool to develop a model and tree which would help us identify the crime trend based on the available data. We approached the design aspect of our data in RapidMiner by retrieving the dataset, and changing column types according to the data within. For the role, we chose the type of offense where the values are either: felony, misdemeanor or violation. We binned the different average temperature values into four temperature ranges. For attributes we used the borough, race, age and sex of suspect and victim to develop a model decision tree. For Decision Tree Criterion we have selected Information\_gain.

We chose the type/category of offense as the Label, which is our target attribute, because we wanted to see what type of crime is being committed and what kind of environmental and individualistic factors are affecting it.

**Model evaluation**

After carefully selecting all the attributes and operators we have created the pipeline which would process the data to generate the results comprising the decision tree, performance vector and example set (validation).

For the first iteration we had selected maximal depth at 4 and ran the model which gave us accuracy of 53.08%.The initial model could be improved so we tried different attributes and used different depths to improve the decision tree and performance vector. By increasing the depth to 10 we got a clear view of information to make business insights and improvements. By making changes to the initial model we achieved accuracy of 55.86%. This produced a more elaborate tree and the model allowed us to find specific crime factors and types based on related parameters. The precision and recall values improved drastically from initial iteration to final iteration. (Appendix 10-11)

# 

# **Scoring the dataset**

The F-measure increased for all types of crime in the final model. After looking closely and understanding the business requirements we recommend high precision as selection criteria for the model. We want lower false positives as that would mean we have identified the correct type of crime and we can pursue the legal action accordingly. Charging a criminal for a higher crime level than what he/she committed would not be ethically right.

Recall in this model takes a secondary step as it would mean we did not acknowledge the level of crime. The costs and efforts involved in identifying the level of crime are high but identifying the correct type of crime to punish the criminal is more important. As a business decision our primary focus would be on precision. We rather be prudent than incarcerate someone for a small crime.

**Business Insights and Recommendations**

Predicting the correct type of crime will lead to better formulation of policies and safeguards. With this data we can better advice government policy and affect changes in crime related policies. By accurately predicting different types of crime, we can better identify useful policy and recommend appropriate punishment that correlates with a specific crime level. Furthermore, we can ensure that NYPD resources are efficiently allocated. Understanding the true data for a crime can also help us advise advocacy groups and other agencies regarding how best to allocate their own resources. While all incidents mentioned in our data set correspond to criminal reports not all crimes are equal.

we were not able to find any direct trends that would allow us to predict the type of crimes related to temperature as it was not linear regression but we got nonlinear bell curved regression analysis results.. The conclusion we come to is that there is not enough data or correlation to confidently predict crime in New York City.

We would recommend expanding the scope to add at least five years of individual yearly and combined analysis. Any useful analysis would also benefit from the addition of judiciary outputs, such as verdict, length of time served, and other related categories. Furthermore, it would be preferable to refer to government sourced data that can ensure a high level of accuracy which could be utilized for accurate analysis and prediction to reduce the crimes.

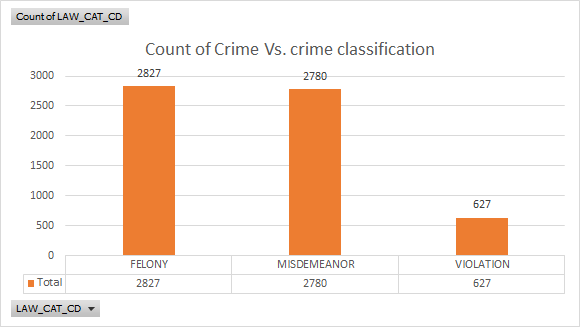
**Data Limitations and Ethical Issues**

There is an inherent danger in using a new tool that is not well understood, does not develop beyond existing social problems, and reinforces embedded negative stereotypes. When looking at the data it is important to err on the side of caution. It is preferable to let one man free for a past misdemeanor rather than incarcerate an innocent person. In addition, we must take into consideration the nature of predictive data. In reality, a past crime may predict the probability of a future crime, but it can not know the future. The unique human element associated with crime is a major variable that cannot be overlooked. While a computer can be programmed to identify behavioral patterns, it cannot be relied on to accurately identify a future human behavior.

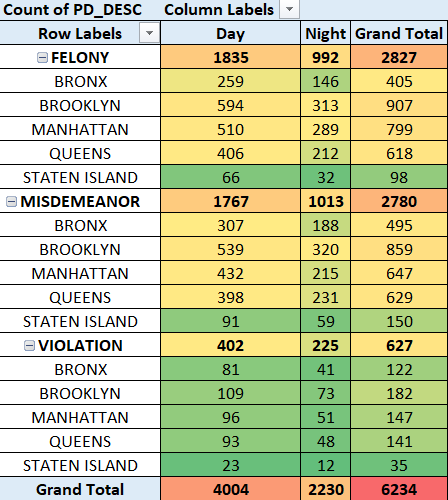
Predictive analytics cannot take into account the circumstances in which a crime is committed and is therefore constrained to deontological ethics. A set of rules can be programmed into Rapid Miner and used to judge an action as right or wrong. The program will then provide suggestive responses based on previous actions and trends. It is important that we be extra cautious, since Rapid Miner relies on data points and cannot understand the situational occurrences.

Although our goal is to reduce crimes by identifying the patterns related to various levels of offense there are a number of limitations associated with this data. If utilizing this method for practical, real time, scenarios.

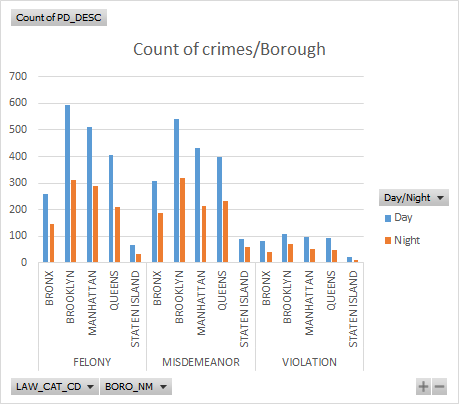
Appendix 1: Count of Crimes Vs. Classification

****

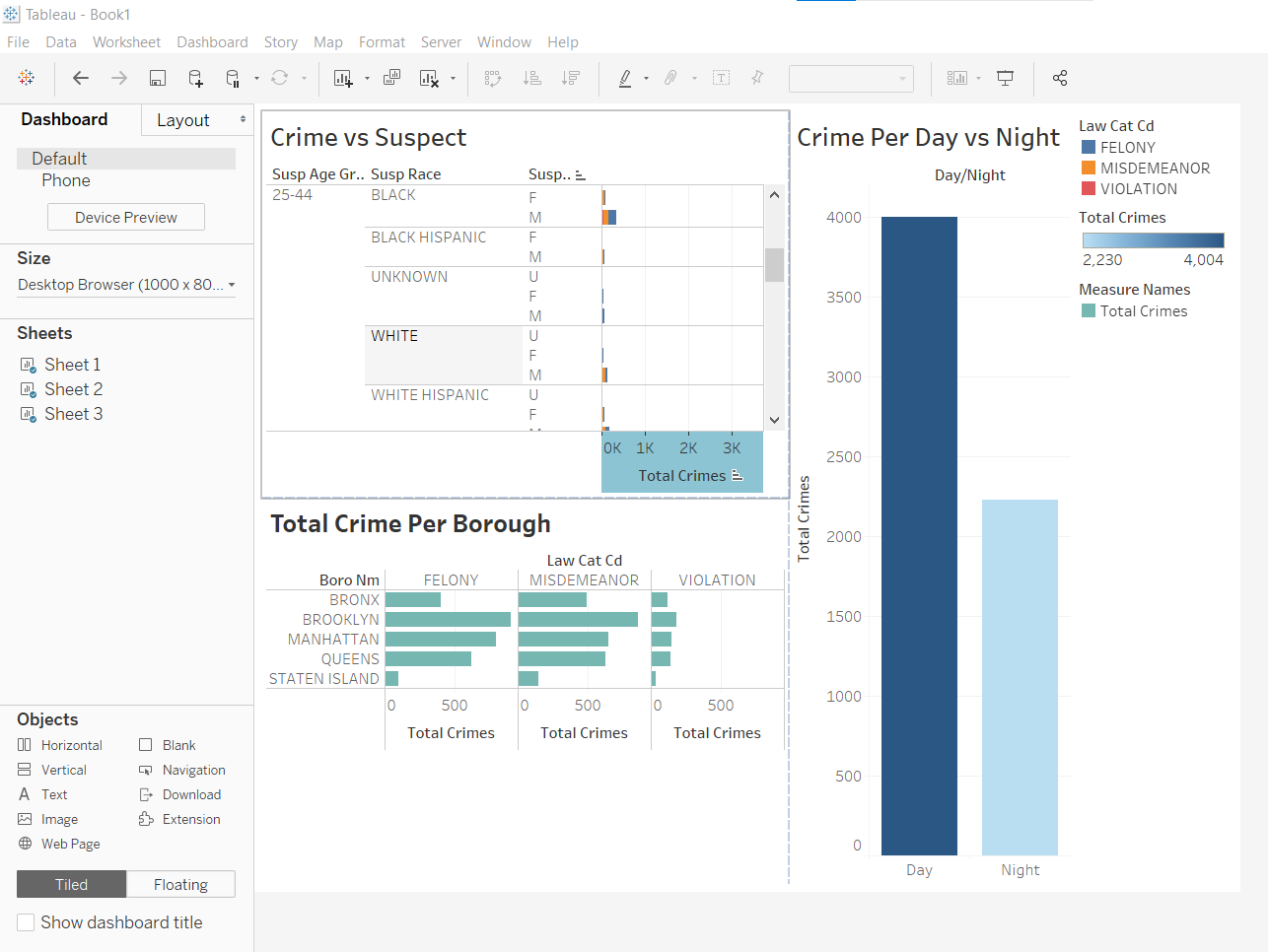
Appendix 2: Pivot of Data for understanding

****

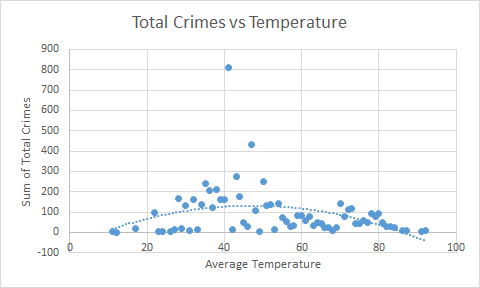
Appendix 3: Count of crimes per Borough

****

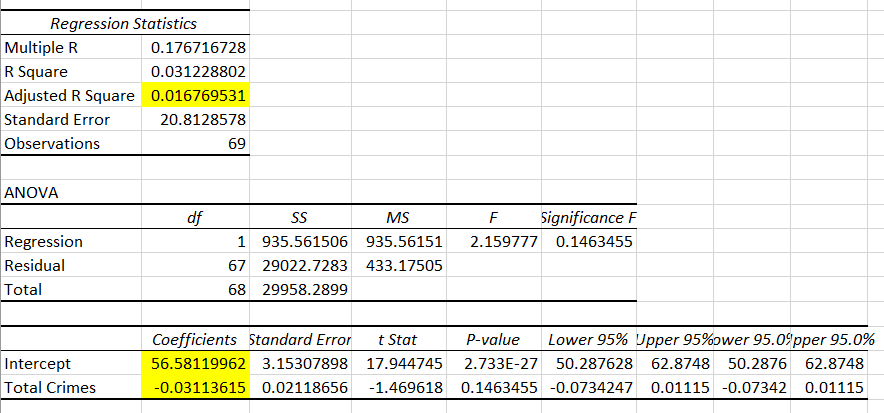
Appendix 4 : Tableau Dashboard

****

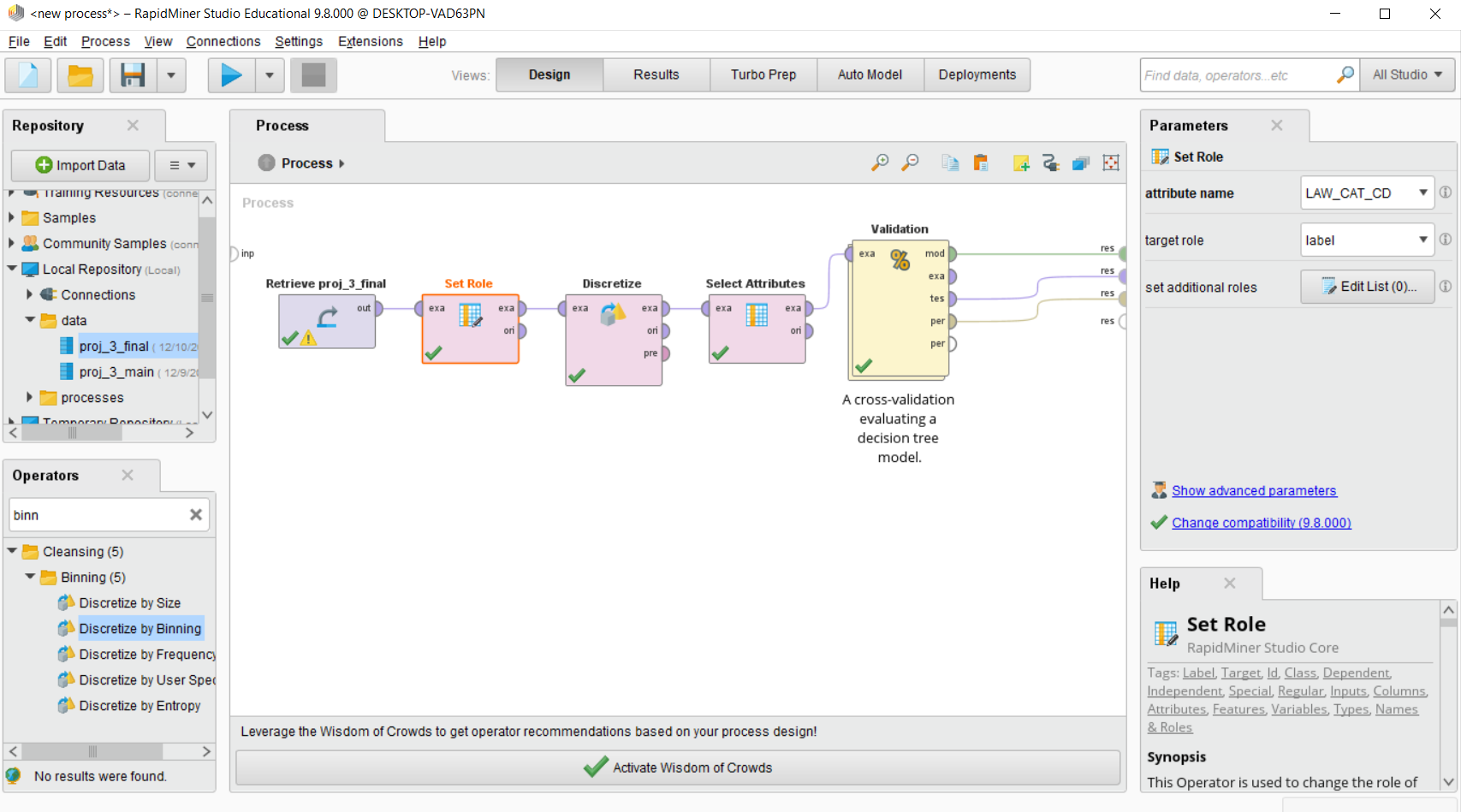
Appendix 5: Total Crimes vs Temperature

****

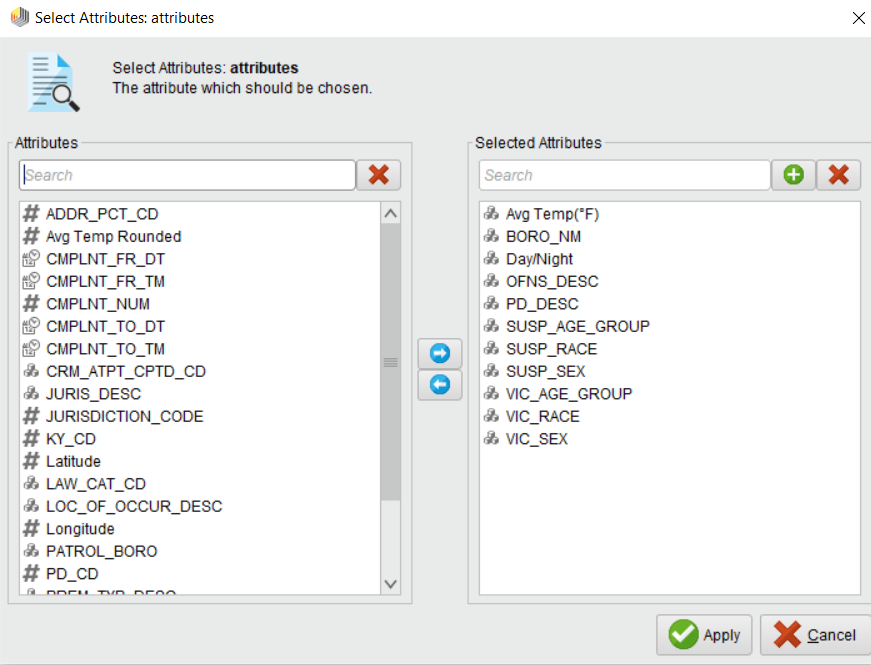
Appendix 6: Regression Analysis

****

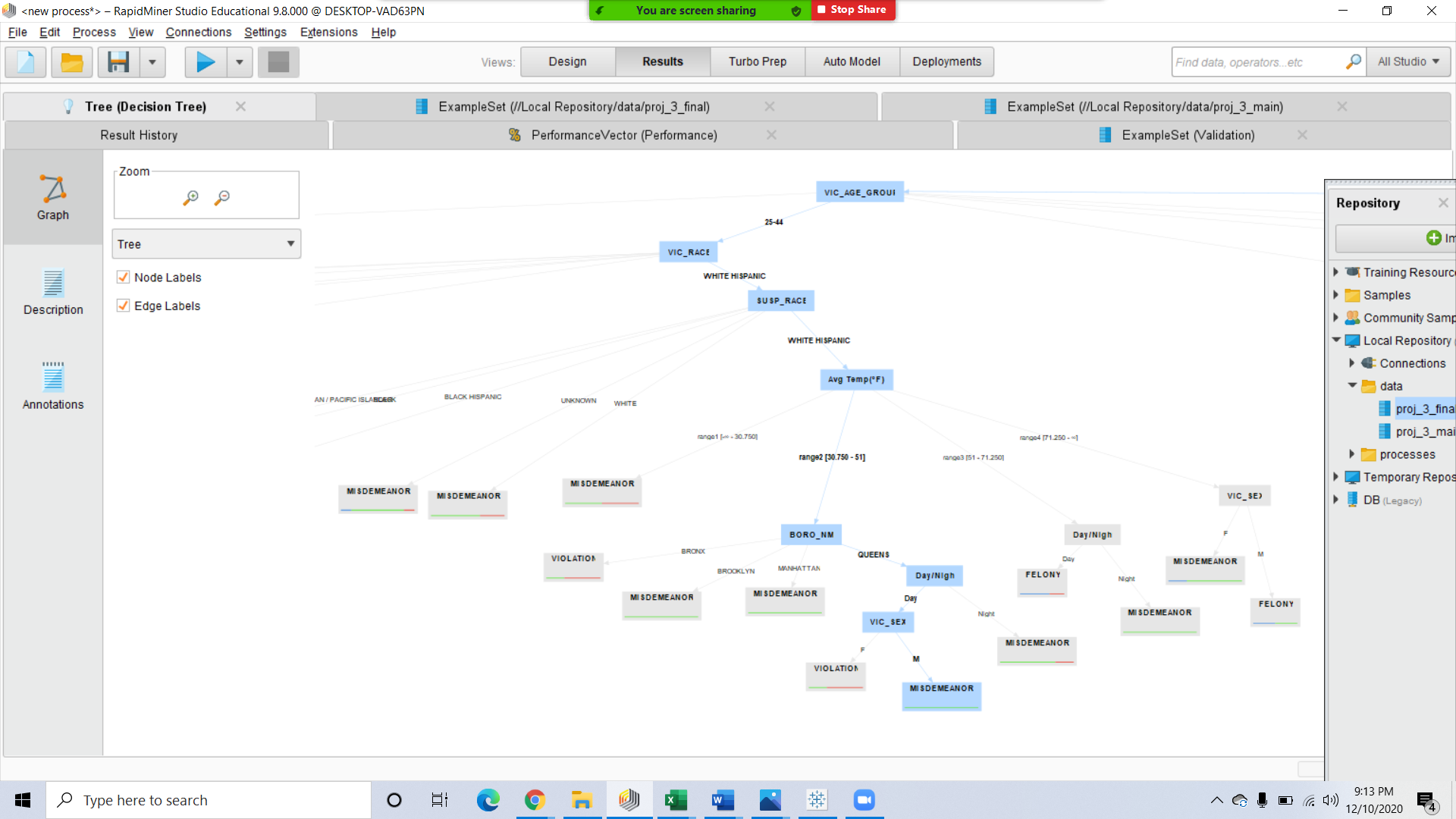
Appendix 7: RAPIDMIINER Process flow

****

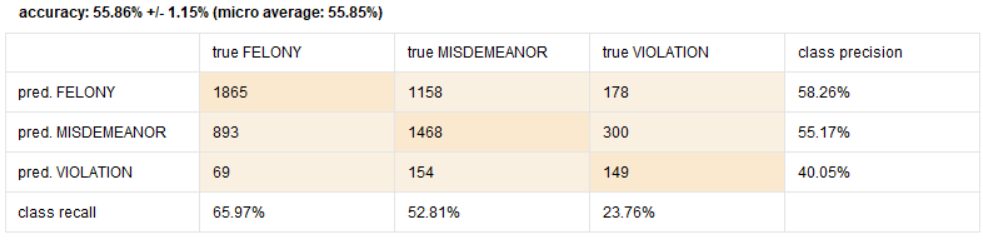
Appendix 8: Attribute Selection

****

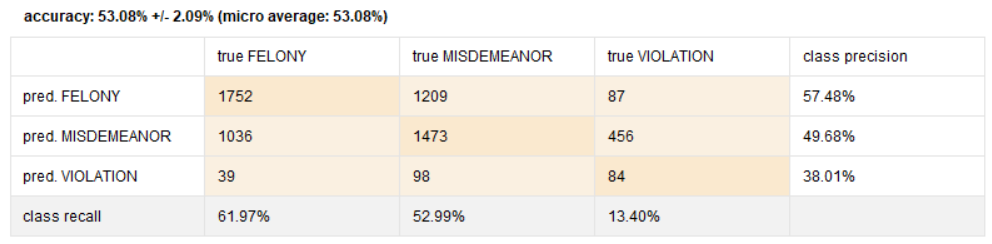
Appendix 9: Final Model

****

APPENDIX 10 :First Confusion Matrix

****

APPENDIX 11: Final Confusion Matrix

****