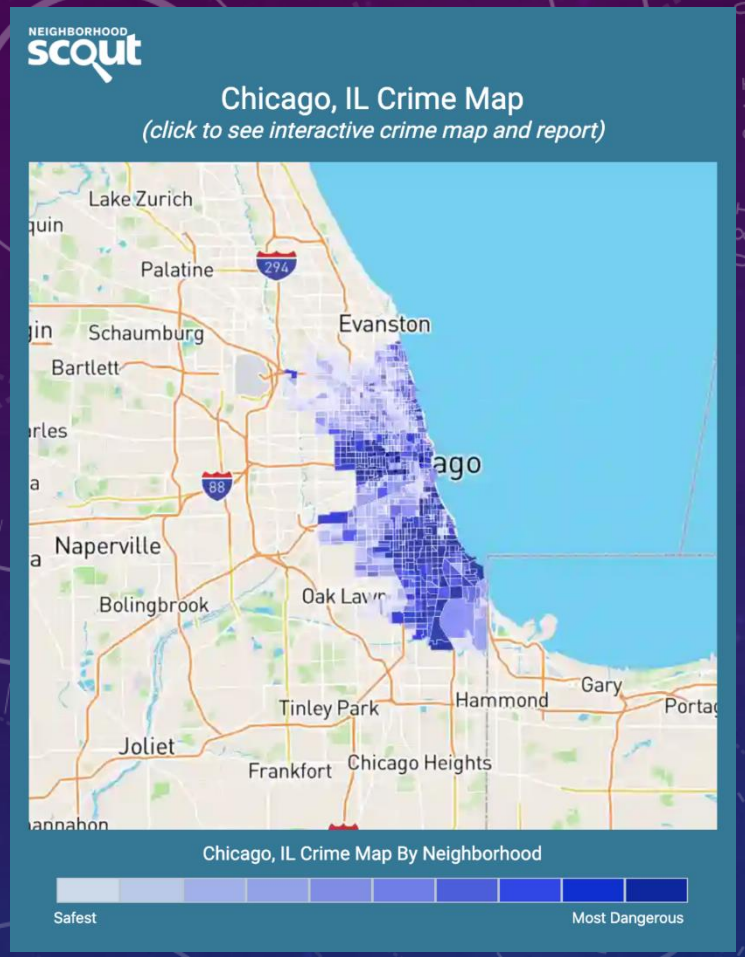


PROJECT 4: PREDICTING TYPE OF CRIME/INCIDENT IN CHICAGO

ANKIT GUBILIGARI

CSE 180

QUESTIONS/OVERVIEW OF ANALYSIS:

- Can I predict the type of crime committed from the numerical and categorical variables given by the dataset?
- Are the predictions of a single decision tree better or worse than the predictions of an ensemble?
- What columns of the table are the most influential in my BigML prediction?
- Which model will perform more efficiently at predicting category of crime?
- How should the city use this analysis to react to this problem?

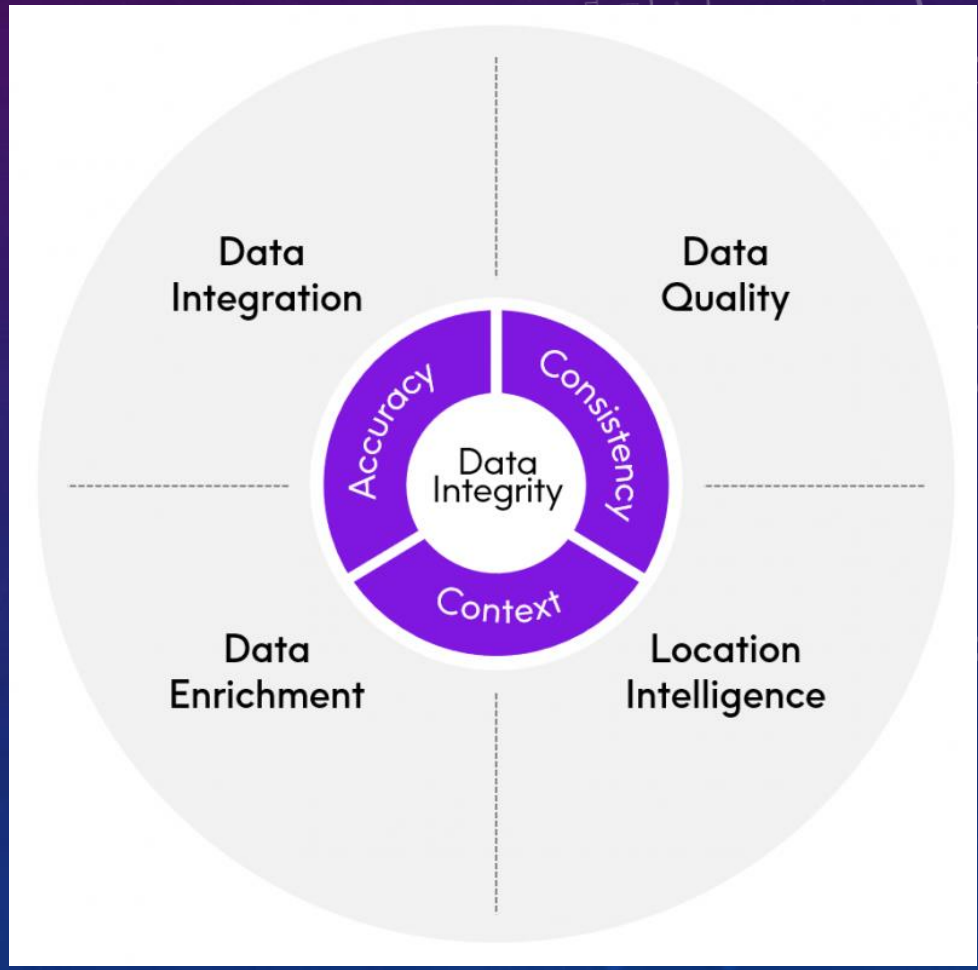
DESCRIPTION OF DATA SOURCE AND BACKGROUND:

- Type: Official
- Source: http://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e (publically available)
- Only shows data from 2001-2017
- 490 MB (medium-sized data pool)

- Background: Contains case #, date, primary type of crime, description of crime, location, arrest status, domestic status, district, ward, year, and location description. This data set classifies the crimes as Battery, theft, assault, motor vehicle theft, and many more.

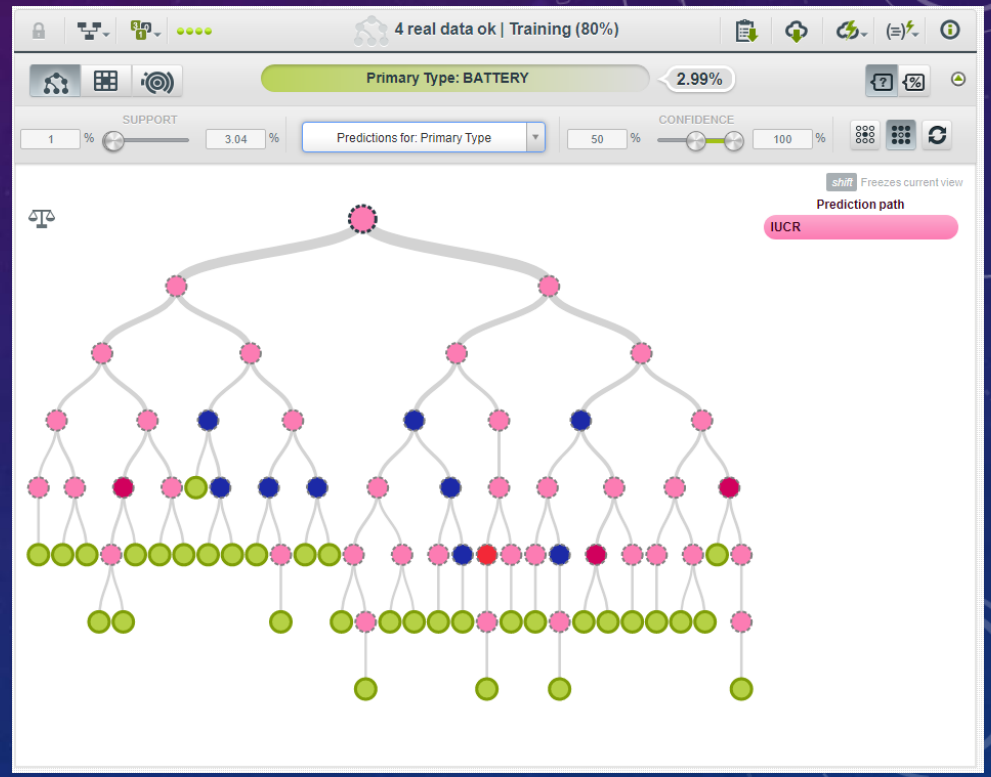
DATA INTEGRITY EVALUATIONS:

- Appropriate Collection Methods: Medium
- Collection Ethics: Medium
- Source Credibility: High
- Appropriate Provenance and Curation: Low
- Data Bias: N/A
- Source Cleanliness: High (Everything was filled in)
- Data Fairly Collected: Medium

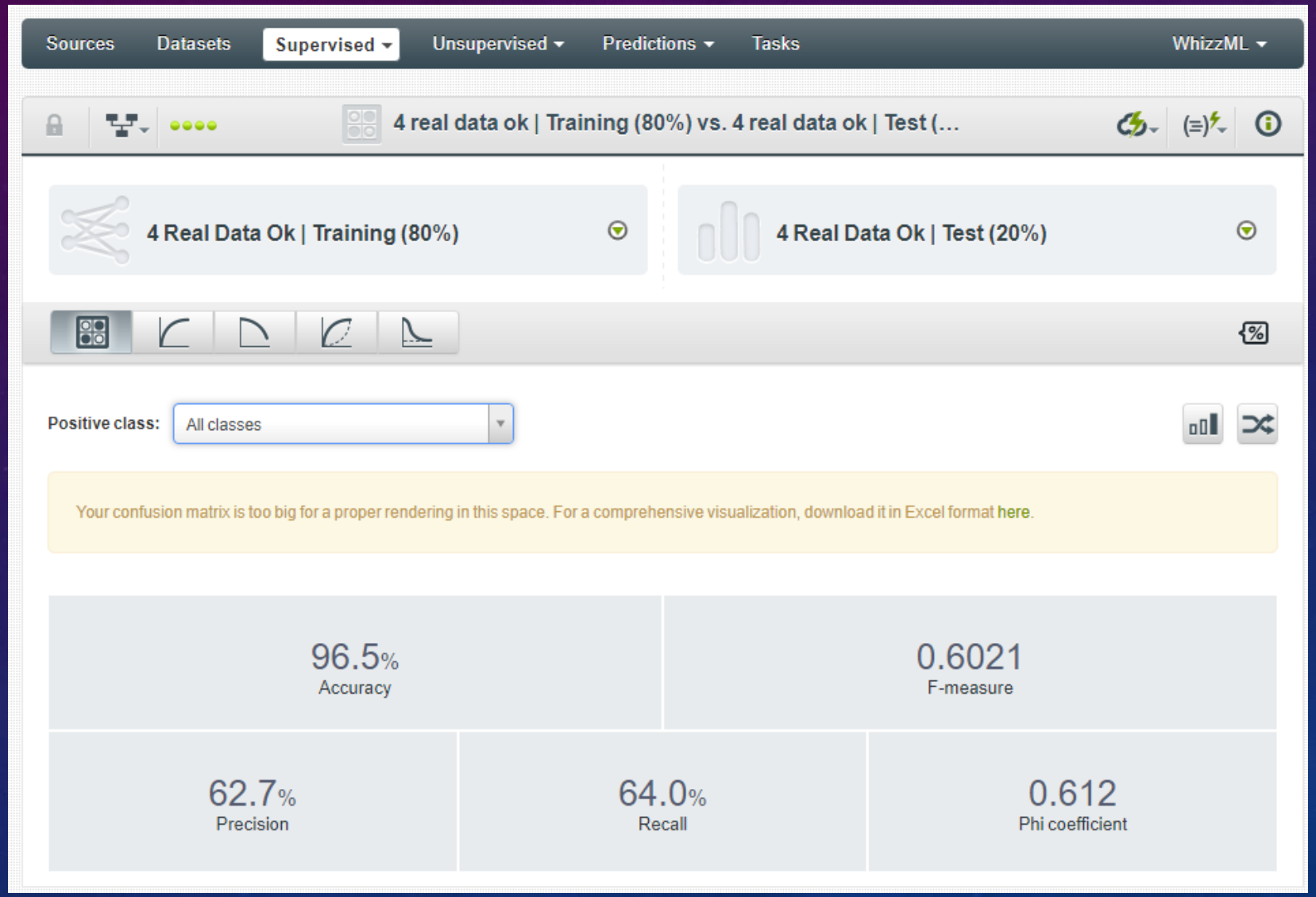


DATA PREPARATION:

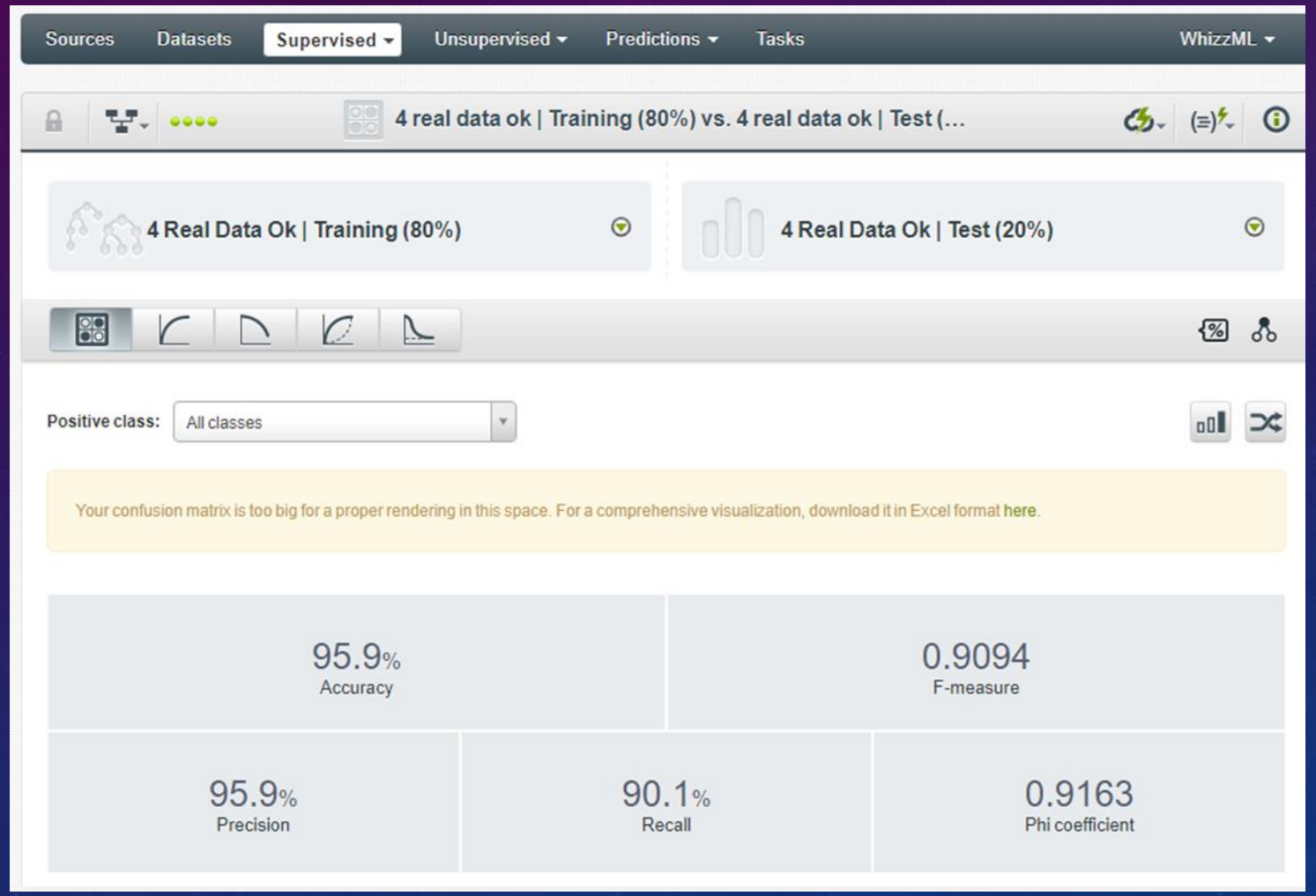
- To the right is a decision tree (single) for all the variables that were included and from this data preparation, I have concluded that I will either use ensemble or deepnet to make my predictions due to the fact that they are more accurate/precise.
- Change certain variables from text to a value in order to have BigMl interpret the data correctly.
- Removed certain rows and columns that were unnecessary or would cause conflict with BigML modelling software
    - These included things like the location description and the identification codes attached to the crimes.



ENSEMBLE EVALUATION FOR TRAINING VS. TEST



DEEPNET EVALUATION FOR TRAINING VS. TEST



WHICH SHOULD I BASE MY PREDICTIONS ON?

- Ensemble: High accuracy, medium F-measure, medium precision, medium and recall
- Deepnet: Higher accuracy, F-measure, precision, and recall

- I will use deepnet to make my predictions.

OBSERVATIONS AND CAVEATS:

- Observations:
  - The deepnet has a 0.6% higher accuracy than the ensemble.
  - The deepnet has about a 0.3 higher f-measure value.
  - The deepnet has a high precision of 95.9% while the ensemble has a medium precision of 62.7%.
  - The deepnet has a 26.1% higher recall value than the ensemble.

- Conclusion: Use deepnet for the prediction to maximize efficiency with its high precision and high recall.

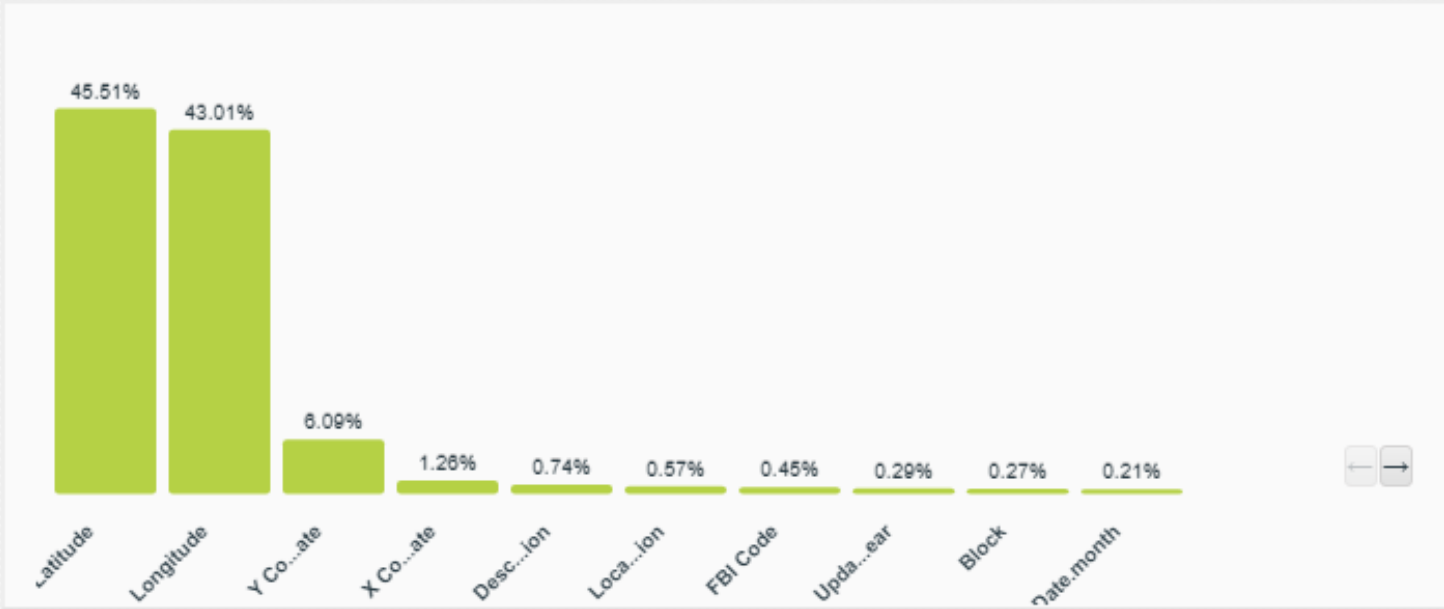
- Caveats:
  - The caveats of using a ensemble are that these medium sized values for the precision, f-measure, and recall rate could lead to inaccurate results that may be skewed.
  - I could not get the confusion matrix to display on the screen without opening excel.
  - Dataset was very large and had a lot of errors I had to clean up at first.

Deepnet Summary Report

×

Field importance

PNG CSV JSON



Close

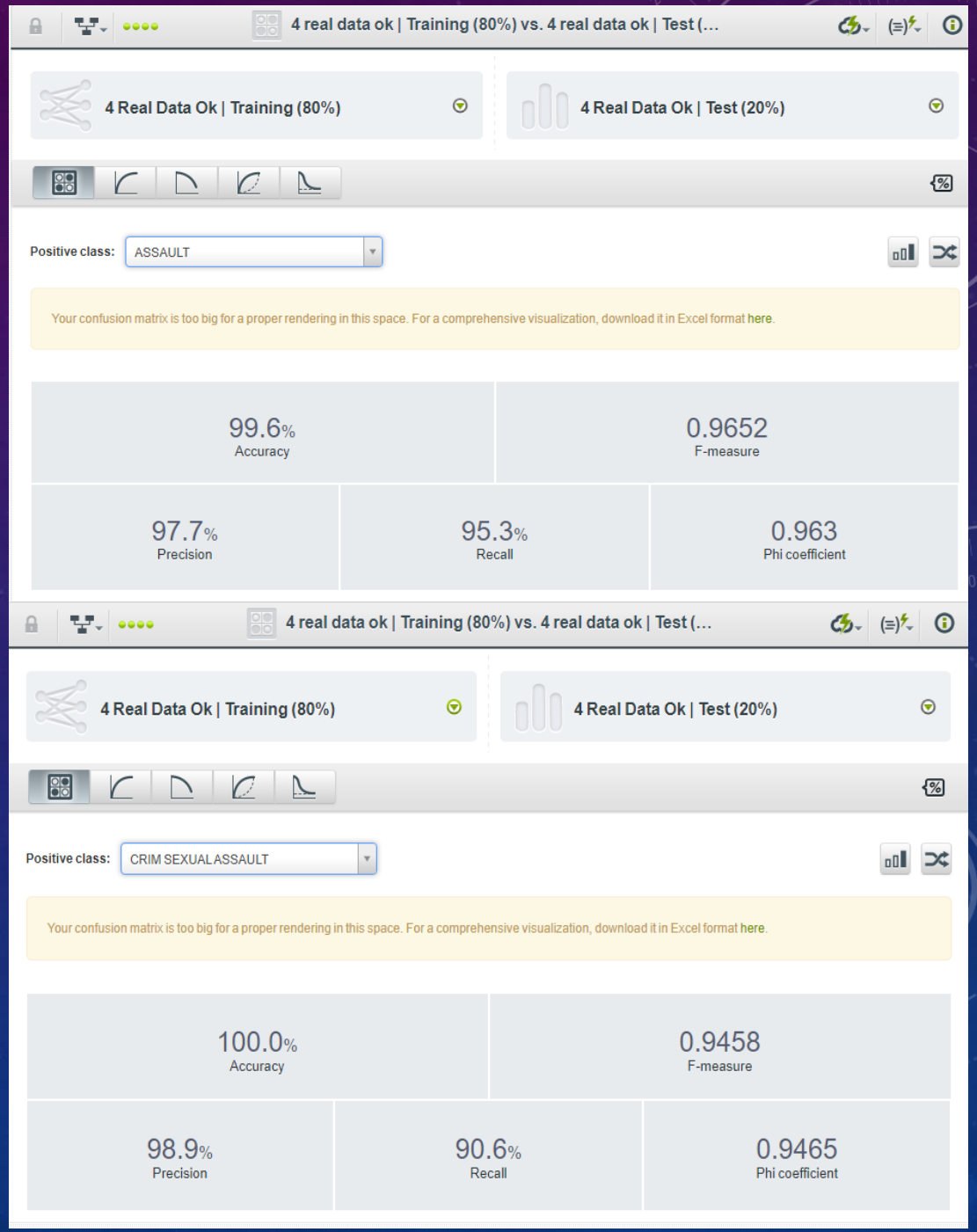
OBSERVATIONS AND CAVEATS:

- Observations:
  - The columns of the table that were the most influential in making the prediction as shown in the slide previously was latitude/longitude followed by description, location, and FBI code.
  - All of the remaining fields can be used but they all have a very small role in the prediction creation.
  - The FBI code was an extremely helpful variable that had nothing to do with geographical position.

- Caveats:
  - It could not fit all of the categories/columns into the display and you have to manually scroll through all of them.
  - Since the location in latitude and longitude is the highest influence on the prediction, this analysis and dataset may be very poor quality as the other columns should have a higher impact.

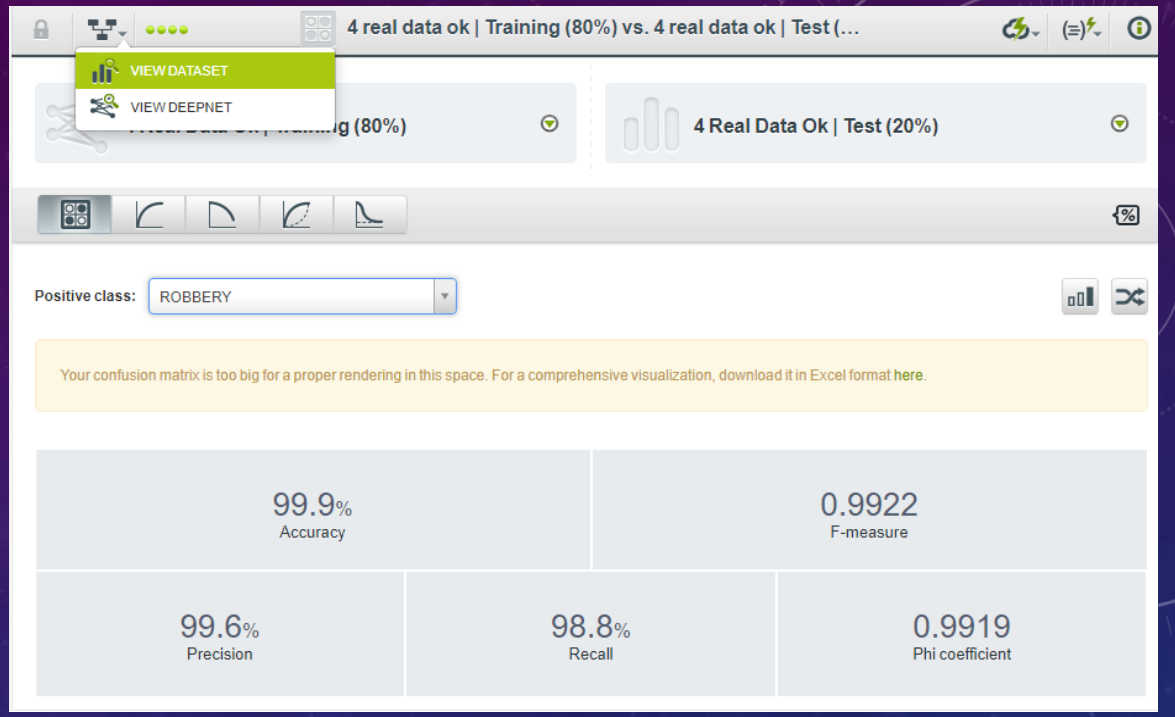
- The deepnet for assault has very high accuracy, f-measure, precision and relatively high recall %.

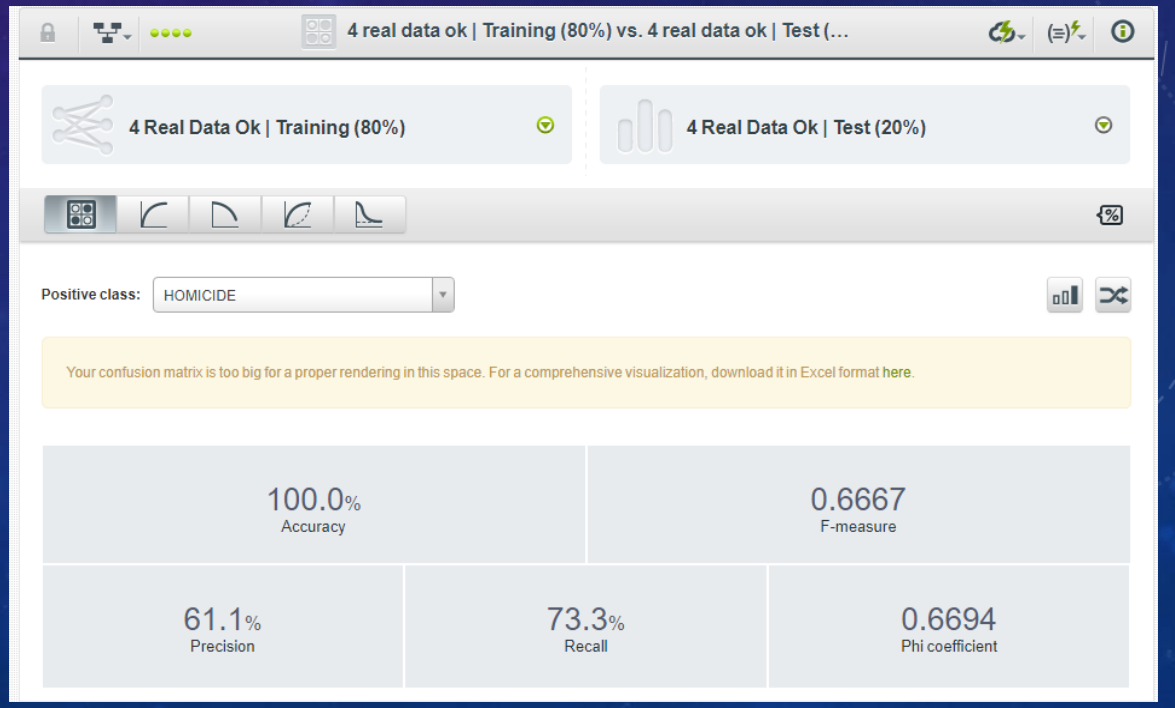
- The deepnet for Criminal Sexual Assault has 100% accuracy, relatively high f-measure, high precision and relatively high recall %.



- The deepnet for robbery has very high accuracy, very high f-measure, high precision and high recall %.

- The deepnet for Criminal Sexual Assault has 100% accuracy, medium f-measure, bad precision and medium recall %.





OBSERVATIONS AND CAVEATS:

- Observations:
  - Certain types of crimes had an extremely perfect model prediction such as robbery and criminal sexual assault just to name a few. These crimes had a near 100% accuracy, very high precision measure, and a high f-measure.
  - Most of the different types of crimes had the model performing very well however there are some outliers with a very average f-measure value.
  - One thing I noticed was that nothing really fell below 95% accuracy and 90% precision throughout all of the data points in the prediction.

- Caveats:
  - I could not present all of the different types of crime predictions within this powerpoint because of the magnitude of types of crimes present within this data set.
  - All of the types of crimes being modeled so perfectly could indicate that this was a subpar data set or something went wrong.
  - Some of the types of crimes did not contain f-measure values and recall values.

CONCLUSION FOR HOW CHICAGO SHOULD USE THIS ANALYSIS:

- This deepnet model prediction is presenting accurate results; however, the city of Chicago needs to be extremely careful when using this analysis to institute reforms.
- The extremely high levels of accuracy, precision, f-measures, and recall % for all of the types of crimes committed indicates that there is something else that is providing this effect on the prediction by BigML.
- I would recommend that the city of Chicago creates a new data set without certain data points and variables such as location that could help pinpoint the other variables that can help predict which type of crime has occurred.
- Once a new cleaned up data set has been created, the city of Chicago should do their own analysis with either deepnet or ensemble in order to determine the most influential factors on the type of crime.
- For this analysis, Deepnet provided a lot of helpful insight because of the high rate of precision over the recall rate.