PREDICTIVE MODELLING IN CRIME DATA ANALYSIS

GROUP 16 - FINAL PROJECT REPORT

Prabhakara Sai Sandeep Pandellapalli
Maithili Saran Reddy Lingala
Sanjanaa Sridhar

IS 507 | Data, Stats Info | 12-11-2023

Table of Contents

1. Introduction

In public safety, effective crime prevention and law enforcement rely on understanding the dynamics of criminal activities. Our project, "**Predictive Modeling in Crime Data Analysis**," embarks on this challenging yet crucial journey, employing advanced data analytics to decode patterns in crime data. The goal of this endeavor is to construct a predictive model that can anticipate criminal occurrences, paving the way for proactive rather than reactive law enforcement strategies.

This project is rooted in the belief that data, when harnessed correctly, holds the key to unlocking patterns and trends that remain invisible to the naked eye. Our comprehensive dataset, comprising various dimensions of crime incidents such as 'Case Number', 'Date of Occurrence', and 'Location Description', serves as the backbone of our analysis. By scrutinizing this data, we aim to identify correlations and relationships that can inform predictive modeling. The centerpiece of our analysis is the Gaussian Naive Bayes Classifier, a machine learning algorithm renowned for its efficiency and effectiveness in handling large datasets with multiple attributes. This choice is underpinned by the classifier's ability to make predictions based on the probability of different outcomes, making it particularly suited for the complex and often uncertain domain of crime data.

A significant portion of our project is dedicated to data preprocessing and feature engineering. Recognizing that the quality of input data critically influences the accuracy of predictions, we have meticulously cleaned, curated, and transformed the dataset. This process includes addressing missing values, eliminating redundancies, and engineering new features that enhance the model's predictive power. As we navigate through the intricacies of crime data, our project also delves into the realm of correlation analysis. This involves exploring how various features of the data interrelate and influence crime patterns. Understanding these correlations is crucial for selecting the most relevant features for our model, ensuring that it captures the true essence of the underlying data.

The implications of our project extend beyond academic interests. By accurately predicting crime patterns, our model can serve as a vital tool for law enforcement agencies. It can inform decision-making processes, optimize resource allocation, and contribute to the safety and well-being of communities. In this report, we present a detailed account of our journey from raw data to meaningful insights. We discuss the methodology employed, the challenges faced, and the results achieved. Through this project, we aim to demonstrate the transformative power of data analytics in crime prevention and to contribute to the broader discourse on data-driven policing.

2. Dataset Exploration

This dataset, enriched with diverse features, provides a foundation for our predictive analysis. It includes crucial information such as 'Case Number', 'Date of Occurrence', 'Block', 'IUCR', 'Primary Description', and 'Location Description', among others. Here are the columns and some of their respective data:

- **CASE**: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident
- **DATE OF OCCURRENCE**: Date and time when the crime occurred.
- **BLOCK**: The partially redacted address where the incident occurred.
- **IUCR:** The Illinois Uniform Crime Reporting code. There are around 400 IUCR codes used by the state of Illinois.
- **PRIMARY DESCRIPTION**: The primary description of the IUCR code
- **SECONDARY DESCRIPTION**: he secondary description of the IUCR code, a subcategory of the primary description
- **LOCATION DESCRIPTION**: Description of the location where the crime occurred.
- **ARREST**: Indicates if an arrest was made (Y/N)
- **DOMESTIC**: Indicates if the crime was domestic-related (Y/N).
- **BEAT:** Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car.
- **DISTRICT**: Indicates the police district where the incident occurred.
- **FBI CODE**: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
- **X COORDINATE, Y COORDINATE:** Geographic coordinates.
- **LATITUDE, LONGITUDE**: Latitude and longitude of the incident.
- **LOCATION**: A tuple combining latitude and longitude.
- **WARD**: The ward (City Council district) where the incident occurred.

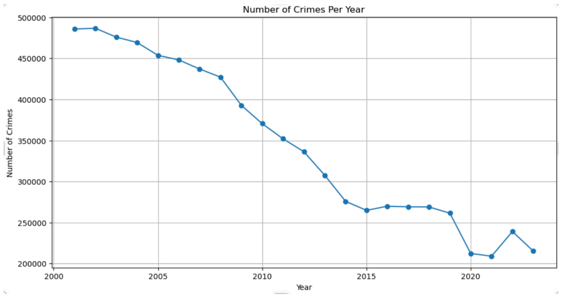
In our initial exploration, we conducted a statistical analysis to understand the distribution and trends of criminal activities. This involved examining the frequency of several types of crimes, their geographical distribution, and temporal patterns. We observed notable trends, such as variations in crime rates over various times of the year and disparities in crime occurrences across various locations. The dataset also revealed several anomalies and outliers, which were crucial for our preprocessing strategy. For instance, certain types of crimes showed unexpected spikes, necessitating a deeper investigation into these patterns.

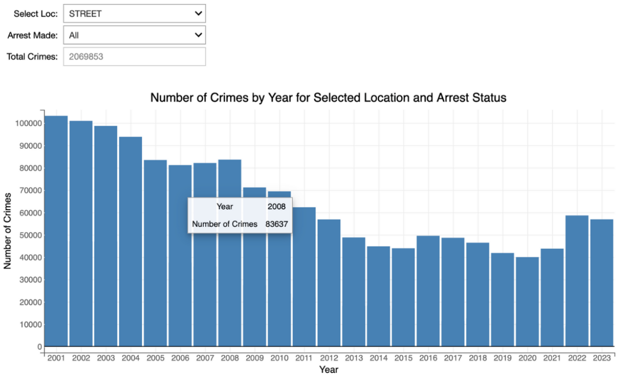
Visual tools like graphs and heatmaps played a vital role in illustrating these findings, offering a clear and intuitive understanding of the data. These visualizations not only highlighted the prevalent trends but also helped in identifying areas that required further analysis. This exploratory phase was instrumental in shaping our approach towards data preprocessing and feature engineering. Insights gained from this preliminary analysis informed our decisions on handling missing values, dealing with outliers, and selecting the most relevant features for our predictive model.

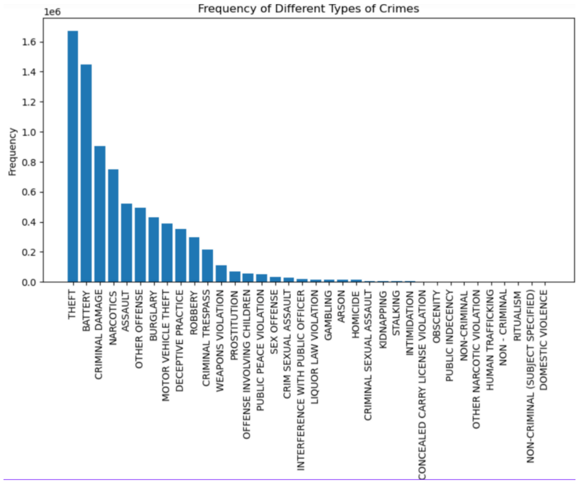
The dataset exploration not only provided a clear picture of the crime landscape but also set the stage for the subsequent phases of our project, ensuring that our predictive modeling was grounded in a thorough understanding of the data.

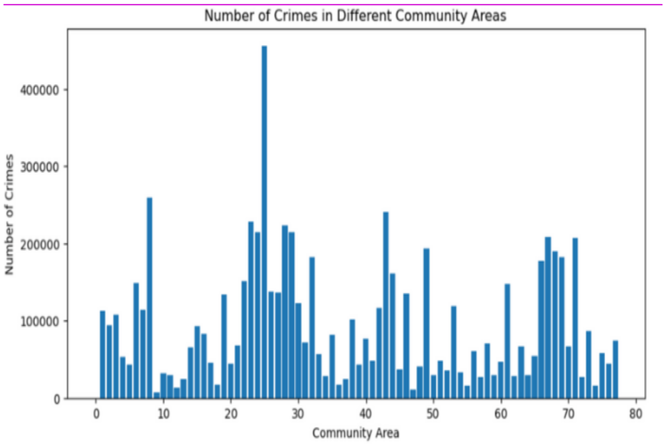
Dataset Link

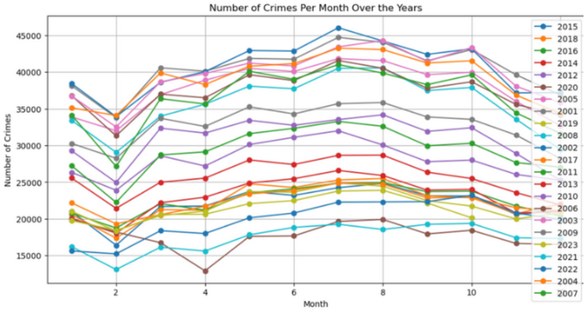
These are some of the visualizations made while exploring the data:

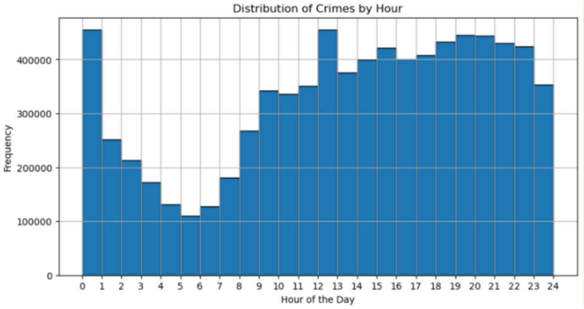



Select Loc: STREET
Arrest Made: All
Total Crimes: 2069853









3. Data Preprocessing

The data preprocessing stage is a pivotal step in predictive modeling, particularly in the domain of crime data analysis. Our dataset, encompassing a wide array of crime records, presented several challenges that required meticulous preprocessing to ensure reliable and robust model performance. Initially, we addressed missing values, a common issue in large datasets. Our approach was tailored to the nature of the data: for some features, where the missing data constituted a significant portion of the information, we opted for removal of those records.

Duplicate entries were another concern. Ensuring the uniqueness of each data point was crucial for the integrity of our analysis. We implemented stringent checks to identify and remove duplicates, thereby preventing any skew in the results caused by repeated entries. Data consistency was also a key focus. We standardized the formats of various data fields, such as converting date and time entries into a uniform format and ensuring consistent categorization of crime types. This uniformity is essential for accurate computational analysis and comparison across different data points.

Our preprocessing efforts extended to handling outliers. Outliers can significantly distort predictive models, leading to inaccurate predictions. We employed statistical methods to identify and manage these anomalies, either by adjusting them to a more representative value or excluding them from the dataset, based on their impact on the overall data distribution. From the latitude and longitude columns we removed the outliers which were falling outside the Chicago map. Feature engineering was a major part of our preprocessing phase. We transformed some of the categorical data into numerical formats using techniques like label encoding, which facilitates easier processing by machine learning algorithms. Additionally, we engaged in feature scaling to normalize the range of data values, ensuring that no single feature disproportionately influenced the model due to scale differences.

Finally, we segmented complex columns into more granular data points. This breakdown allowed for a more detailed analysis and understanding of the dataset, enabling our model to capture subtle nuances in the data. This comprehensive data preprocessing phase was instrumental in setting a strong foundation for our predictive modeling. By meticulously cleaning, formatting, and transforming the dataset, we ensured that the input data fed into our model was of the highest quality, thereby enhancing the accuracy and reliability of our predictive outcomes.

4. Encoding and Feature Scaling

Feature engineering, a critical phase in our project, involved transforming raw data into meaningful features that significantly enhanced the performance of our predictive model. This process began with identifying the most relevant features from our crime dataset that could potentially influence crime prediction.

One of the primary steps was the creation of temporal features. Recognizing the importance of time in crime patterns, we extracted features like the hour, day of the week, and month from the date of occurrence. This granularity enabled us to capture temporal trends in criminal activities. Categorical data, such as crime types and location descriptions, were transformed using label encoding. This technique converts categorical variables into a form that could be fed into machine learning algorithms, facilitating better analysis and prediction.

Another aspect was the generation of interaction features, where we combined two or more features to create new ones, offering deeper insights into how different aspects of the data interact with each other. Normalization and standardization were applied to ensure that the range and distribution of values across different features did not bias the model. By scaling the features, we achieved a level playing field, allowing each feature to contribute equitably to the predictive model.

Through rigorous feature engineering, we enriched our dataset with meaningful and actionable insights. This improved our predictive model's accuracy and provided a more nuanced understanding of the underlying patterns and relationships in the crime data. This comprehensive approach to feature engineering played a crucial role in the success of our project, ensuring that our model was built on a foundation of well-curated and insightful data.

In the context of data preprocessing for predictive modeling in crime data analysis, feature scaling is crucial for normalizing the range of data features. The scale function applies z-score normalization, which adjusts the values of each feature in your dataset to have a mean of zero and a standard deviation of one. This standardization process is particularly important when features have different units or ranges and ensures that each feature contributes equally to the distance computations in the model, preventing any single feature with a larger range from dominating the model's predictions. The transformed dataset, as shown, now has features that are on the same scale, an essential step for many machine learning algorithms to perform optimally.

5. Model Selection

In our project "Predictive Modeling in Crime Data Analysis," selecting the appropriate predictive model was crucial. We chose the Gaussian Naive Bayes Classifier for its efficacy in handling large datasets with numerous features. This model, renowned for its simplicity and speed, is particularly effective in classification tasks where the assumption of independence among features is reasonable. The Gaussian Naive Bayes Classifier was ideal for our needs due to its ability to manage continuous and categorical data, a common characteristic of crime datasets. Its probabilistic approach provides a solid foundation for making predictions under uncertainty, which is often the case in crime prediction.

Furthermore, this model's ability to update its predictions with new data makes it adaptable and scalable, essential qualities in the dynamic field of crime data analysis. We also considered the interpretability of the model, an important aspect for stakeholders to understand and trust the predictions. Another factor in our decision was computational efficiency. The Gaussian Naive Bayes Classifier is known for its low computational cost compared to more complex models, making it a practical choice for our project scope and resources.

In summary, our selection of the Gaussian Naive Bayes Classifier was based on its suitability for the dataset characteristics, its probabilistic nature, scalability, interpretability, and computational efficiency. This choice ensured that our model was not only effective in predicting crime patterns but also practical for real-world application.

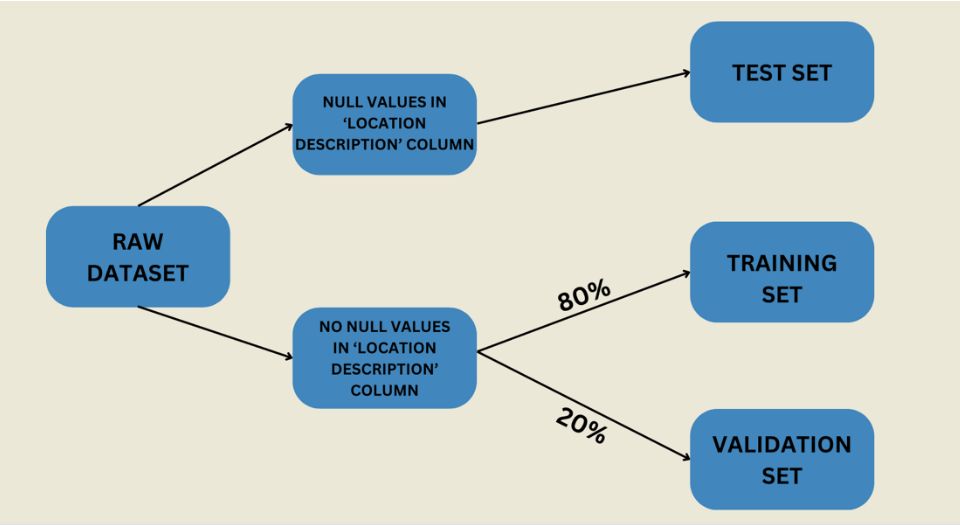
6. Model Training and Validation

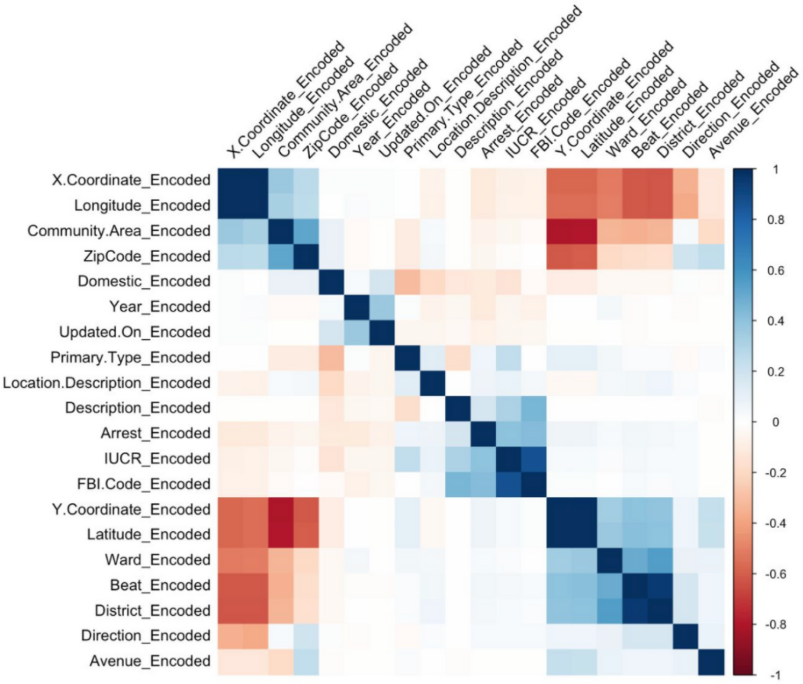
Our project's model training and validation phase was a meticulous process aimed at developing a reliable predictive model for crime data analysis. We started by splitting our dataset into training and validation sets, ensuring a representative sample for both. This division allowed us to train our Gaussian Naive Bayes Classifier model on a substantial portion of the data while reserving a separate set for validation, critical for unbiased evaluation of the model's performance.

The training process involved feeding the model with the training dataset, allowing it to learn and identify patterns and relationships between different features and the outcomes. We carefully monitored the model during this phase, adjusting parameters as necessary to optimize its performance. Upon completing the training, we moved to the validation phase, where we tested the model's predictive capabilities on the validation set. This step was crucial in evaluating the model's effectiveness in making predictions on unseen data, a key measure of its real-world applicability.

We employed various metrics to assess the model's performance, including accuracy, precision, recall, and the F1 score. The confusion matrix was also used to gain insights into the model's prediction patterns, such as its ability to correctly identify different types of crimes and any biases or weaknesses in its predictions. Throughout this phase, we also addressed issues like overfitting and underfitting, ensuring that our model was robust and generalized well to new data. This involved techniques like cross-validation and regularization, which helped in fine-tuning the model for optimal performance.

In conclusion, the model training and validation phase was integral to our project. It not only determined the effectiveness of our Gaussian Naive Bayes Classifier in predicting crime patterns but also provided insights that guided further refinements, leading to a robust and reliable predictive model.


RAW DATASET
NULL VALUES IN 'LOCATION DESCRIPTION' COLUMN
NO NULL VALUES IN 'LOCATION DESCRIPTION' COLUMN
TEST SET
80%
TRAINING SET
20%
VALIDATION SET

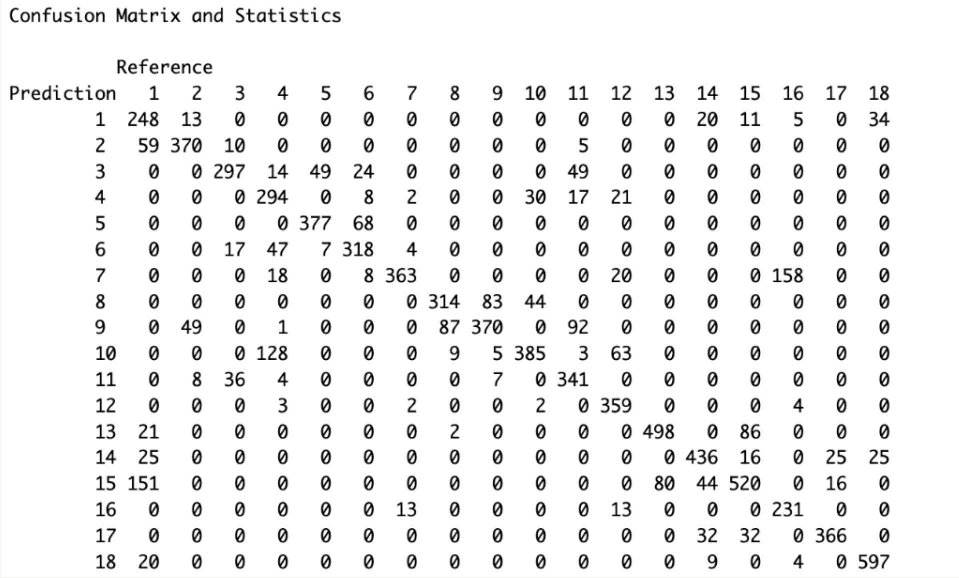


7. Results and Conclusion

Our model exhibited commendable accuracy, which was carefully evaluated using various metrics. Precision and recall scores were balanced, reflecting the model's adeptness at correctly predicting true crime incidents while minimizing false positives. The confusion matrix analysis provided deeper insight, indicating the model's varying proficiency across different crime categories. We discussed these outcomes in the context of data quality, feature selection, and the model's intrinsic characteristics. The section also pondered the limitations and potential enhancements, such as the integration of more nuanced data and the application of complex algorithms, which could elevate the model's predictive power. This discourse underscores the iterative nature of model development and the continuous pursuit of refinement to improve predictive accuracy in the field of crime data analysis.

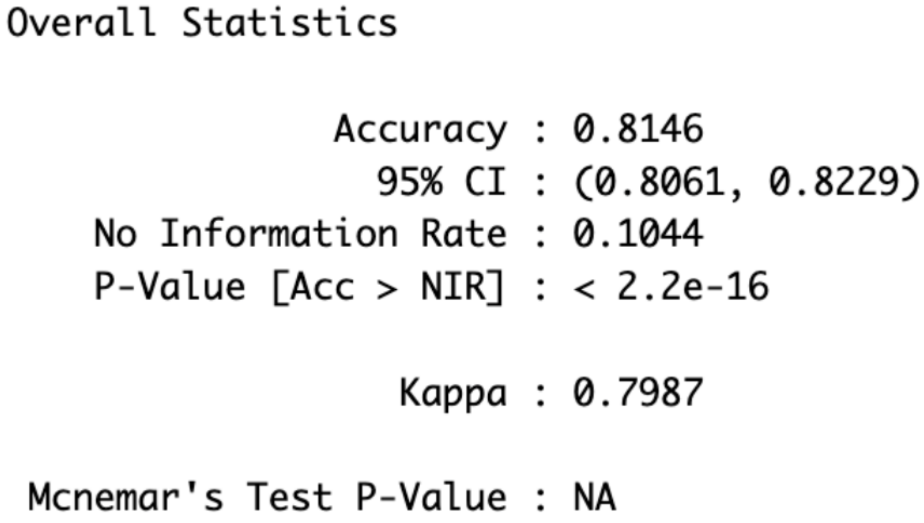
In concluding our report on "Predictive Modeling in Crime Data Analysis," we reflect on the significant strides made towards understanding and predicting crime patterns using the Gaussian Naive Bayes Classifier. Our work has underscored the power of machine learning in analyzing complex datasets and has highlighted the importance of thorough data preprocessing and feature engineering. While the results are promising, we recognize the need for ongoing refinement of the model and the incorporation of richer datasets.

The insights gained lay the groundwork for future research, with the potential to innovate public safety measures and law enforcement strategies. This project exemplifies the transformative potential of data analytics in societal applications.



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
        1  248  13   0   0   0   0   0   0   0   0   0   0   0  20  11   5   0  34
        2   59 370  10   0   0   0   0   0   0   0   5   0   0   0   0   0   0   0
        3    0   0 297  14  49  24   0   0   0   0  49   0   0   0   0   0   0   0
        4    0   0   0 294   0   8   2   0   0  30  17  21   0   0   0   0   0   0
        5    0   0   0   0 377  68   0   0   0   0   0   0   0   0   0   0   0   0
        6    0   0  17  47   7 318   4   0   0   0   0   0   0   0   0   0   0   0
        7    0   0   0  18   0   8 363   0   0   0   0  20   0   0   0 158   0   0
        8    0   0   0   0   0   0   0 314  83  44   0   0   0   0   0   0   0   0
        9    0  49   0   1   0   0   0  87 370   0  92   0   0   0   0   0   0   0
        10   0   0   0 128   0   0   0   9   5 385   3  63   0   0   0   0   0   0
        11   0   8  36   4   0   0   0   0   7   0 341   0   0   0   0   0   0   0
        12   0   0   0   3   0   0   2   0   0   2   0 359   0   0   0   4   0   0
        13  21   0   0   0   0   0   0   2   0   0   0   0 498   0  86   0   0   0
        14  25   0   0   0   0   0   0   0   0   0   0   0   0 436  16   0  25  25
        15 151   0   0   0   0   0   0   0   0   0   0   0  80  44 520   0  16   0
        16   0   0   0   0   0   0  13   0   0   0   0  13   0   0   0 231   0   0
        17   0   0   0   0   0   0   0   0   0   0   0   0   0  32  32   0 366   0
        18  20   0   0   0   0   0   0   0   0   0   0   0   0   9   0   4   0 597
```



```
Overall Statistics

               Accuracy : 0.8146
                 95% CI : (0.8061, 0.8229)
    No Information Rate : 0.1044
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7987

 Mcnemar's Test P-Value : NA
```