2006

Use of data mining for investigation of crime patterns

Manoday D. Padhye
West Virginia University

Use of Data Mining for Investigation of Crime Patterns

by

Manoday D. Padhye

Thesis submitted to The College of Engineering and Mineral Resources at West Virginia University in partial fulfillment of the requirements for the degree of

Master of Science
in
Industrial Engineering

Dr. Rashpal Singh Ahluwalia, Ph.D. (Chair)
Dr. Wafik Iskander, Ph.D.
Dr. Arun Ross, Ph.D.

Department of Industrial and Management Systems Engineering
Morgantown, West Virginia

2006

Keywords: Crime Database, Data Querying, Data Mining, Association Rules, Decision Rules, WEKA

ABSTRACT

Use of Data Mining for Investigation of Crime Patterns

Manoday Dhananjay Padhye

Lot of research is being done to improve the utilization of crime data. This thesis deals with the design and implementation of a crime database and associated search methods to identify crime patterns from the database. The database was created in Microsoft SQL Server (back end). The user interface (front end) and the crime pattern identification software (middle tier) were implemented in ASP.NET. Such a web based approach enables the user to utilize the database from anywhere and at anytime. A general ARFF file can also be generated, for the user in Windows based format to use other Data Mining software such as WEKA for detailed analysis. Further, an effective navigation was provided to make use of the software in a user friendly way.

ACKNOWLEDGEMENT

I thank my advisor Dr. Rashpal Singh Ahluwalia for his continued support, guidance and encouragement during the course of this research work. I also wish to thank my committee members that include Dr. Wafik Iskander and Dr. Arun Ross for their valuable advice and support.

Above all, I wish to thank my colleagues at the research laboratory, friends and specially, my parents for their constant support and blessings for enabling my success and happiness in all my pursuits and endeavors in life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

In recent years, there has been an exponential growth in the amount of data being generated. Unseen to all, such data may seem irrelevant, but to some it is a gold mine that needs to be fathomed for precious information pertaining to the cause. Thus, one is introduced to the terminologies of data generation, data collection, data processing, retention and security. Activities such as purchasing groceries, watching television, traveling, health care, opinion polls, elections, drug purchase, and crime, generate stupendous amounts of data. For example, WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits these data to its massive 7.5 terabyte data storage facility, referred to as the "data warehouse" [1,2]. Previously, such data were lost due to lack of tools. Recent advancements in data collection, storage and manipulation tools, such as phenomenal storage and computational capacity, use of the Internet, advanced surveillance equipments etc., have broadened the scope and limits for the same. Moreover, the increasing dependence on high technology equipment has also eased the process of data collection. For example, each and every credit card that is swiped at the WalMart store generates data that can be valuable not only to WalMart, but to credit card companies, manufacturers, advertisement agencies, financial institutions, etc [3]. Another aspect worth mentioning is the security of the database which has to do with protection not just from losing the data, but also from the data falling into undesirable hands. Also, often the privacy of citizens is breached due to lack of established protocols over the use of data, because the data that look harmless are not necessarily so. For example, with the name and social security number stripped from files, still 87 % of Americans can be identified simply by their date of birth, gender, and five-digit zip code, as established from a review at Carnegie Mellon University [4]. Much to one's relief, several companies are becoming increasingly conscious of this fact.

The data may or may not be in a directly usable form and may need some interpretation based on previous knowledge, experience and most importantly the purpose

of data analysis. This problem is further augmented by sheer volume, texture of the data, and lack of human capability to infer it in different ways. For this reason, many computational tools are used and are broadly termed as "Data Mining Tools." The tools are comprised of basic Statistics and Regression methods, ANOVA, Decision Trees, Rule Based Techniques, and, more importantly, advanced algorithms that use Artificial Intelligence, Neural Networks, etc. Such tools are used to query the database either individually or in a combined form, also called "Hybrid Algorithms" [2]. The result of such data processing is that many important non-obvious relationships can be identified. The applications of Data Mining tools are boundless and basically driven by cost, time constraints, and requirements of the community, business, and the government. The application of Data Mining can be broadly classified in the following fields:

a. Medical and scientific research
b. Commercial and financial institutions
c. Security agencies
d. Sports and entertainment

Applications in medicine can be best exemplified by the fact that there are around 3000 cases of brain tumors each year in United States and almost half of them are fatal. The Children's Memorial Hospital in Chicago is mining the gene expression database for pediatric brain tumors [1]. With this effort, the researchers not only wish to understand the tumors in a better way but also to provide more effective treatment to children. Further, Data Mining has extensive application in space research such as studying stars and their movements, studying weather patterns all over the world, and in understanding electromagnetic bursts. In the commercial field, Data Mining is used by credit card companies primarily to look for patterns of suspicious activities in order to prevent fraud. Stores such as WalMart do Data Mining to determine customer buying patterns and trends.

In an ABC News Broadcast of 15 February 2006, it was reported that the U.S. government is developing a massive computer system called Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement System (ADVISE) that can collect

large amounts of data. It will then search for patterns of terrorist activities by cross-referencing them against U.S. intelligence and law-enforcement records. Moreover, the idea is to identify critical patterns in data that illuminate their motives and intentions. The storage requirements alone are huge -- enough to retain about 1 quadrillion entries. While, privacy concerns have put many hurdles in the way of this program and place some restriction on government use of private data, they do not prevent intelligence agencies from buying information from commercial data collectors [4].

The National Basketball Association (NBA) offers a Data Mining application that can be used in conjunction with video recordings of basketball games to analyze the movements of players which helps coaches orchestrate plays and strategies against their opponents.

There are quite a few Data Mining tools that are available commercially or open-source. Some of these tools are: SAS – Enterprise Miner, JMP, R, TeraData, Clementine, and WEKA. Data Analysis tools such as Microsoft Excel, MiniTab and Statistica can indirectly do Data Mining by searching for vital trends in the data. The primary issue with such software is that they are generalized tools and require significant data preparation and/or coding on the part of the user. Moreover, the data are often not in the form to be interpreted directly. Currently, these tools are being upgraded to reduce user input, to improve cross functionality and to get better results. The main thrust of research is in commercial and health care sector followed by science and technology. It is only recently that Data Mining has gained recognition in the national security field.

1.2 Related Research

Searching databases to extract vital information has been practiced for several years. Pre-9/11 it was driven by commercial needs. Post-9/11 has seen increased use of database search techniques for regional and national security needs. By drawing correlations, Data Mining can lead to identification of crime patterns which will help in crime investigation, and, most importantly, to crime prediction.

The Tucson Police Department was able to trace a narcotics network comprised of approximately 60 criminals using Data Mining tools [5]. It was difficult to detect subgroups, interaction patterns, and the overall structure of the network manually. They utilized clustering and block-modeling methods to identify a chain structure. Various techniques were applied as per their applicability and depending on the investigation. The developed system aided investigators in a better and faster understanding of the network and operations of criminals. Most importantly, it suggested investigative leads to the investigators that otherwise might have been overlooked [6].

Based on past experience, it has been observed that clustering techniques group data items into classes based on characteristics so as to maximize in-class similarity. For example, they can identify suspects those conduct crimes in similar fashion or distinguish among groups belonging to different gangs. While, association rules mine frequent data patterns by treating them as rules, so any discrepancy can be identified as an intrusion. Classification is often used to predict crime trends. It can reduce the time required to identify criminal entities. However, it requires a complete training and testing database. It is also limited by a high degree of missing data values that seem to limit the prediction accuracy.

As per the Regional Crime Analysis Program (RECAP) at Richmond B&E, if one decides to use only one year of crime data for detailed analysis then an analyst must spend 1.5 million minutes, and will be able to perform only 15% of queries and spend approximately 20 minutes of time on each case [7]. Thus, the infeasibility of querying the database for analysis is stressed. Also, it was noted that, to effectively relate crimes, the analyst needs to consider combinations of spatial, demographic, personnel and other data attributes.

In intrusion and fraud detection [8], it was observed that besides scalability and efficiency, the fraud detection task had other problems such as skewed distributions of the training data and non-uniform cost per error, which had not been considered by the Knowledge-Discovery and Data Mining research. Moreover, post pruning succeeded in computing with similar or better fraud detection capabilities, while reducing their size and improving efficiency. Finally, the results of the study clearly demonstrated that

distributed Data Mining techniques that combine multiple models produce more effective fraud and intrusion detectors.

To explore associations among a large number of objects of different types, Link Analysis concepts have been implemented [9]. Used in the case of money laundering, these objects include people, bank accounts, businesses, wire transfers, and cash deposits. Exploring relationships among different objects help indicate networks of activity, both legal and illegal. The technique ensured productive use of records, but proved to be computationally intensive and required great skill and judgment on the part of the analyst during link construction and interpretation of whether the networks represented a legitimate pattern or that of a criminal organization.

The use of statistical methods, predominantly Logistic Regression and Bayesian Networks, has also been suggested [10]. However, these methods are in a primitive stage. There is no clear consensus on how to control groups from the population as a datum for a given type of crime. There is also lack of consensus regarding the selection of a particular method based on crime. Lastly, the interpretation of the results depends on the viewpoint of the statisticians and the crime investigators.

All of the above cases stress the fact that there are many means to the end, and more research is required to efficiently and effectively get the desirable outcomes. In other words, there are many tools at our disposal but not an all-encompassing model or approach [2]. Examples of the current research field reaffirm the fact that, human intervention is absolute. Also, better software and hardware are keys to analyzing the data that are routinely being generated these days.

1.3 Problem Statement

The objective of this thesis is to design and implement a crime database and to provide the user with a capability to detect crime patterns in the data. The data fields will be identified from the crime information in the available literature. Basic and advanced search algorithms will be implemented. Moreover, Association Rules and Decision Rules algorithms will be implemented and validated with the WEKA software.

The database will be created in Microsoft SQL Server (back end). The user interface (front end) and the crime pattern identification software (middle tier) will be implemented in ASP.NET. This will enable the user to utilize the database from anywhere and at anytime. A code to generate ARFF file that is compatible with WEKA will be provided to the user to graphically view the data. Further, an effective navigation will be provided to make use of the software in user friendly way.

It is expected that the Data Mining Tool will be able to identify hidden patterns from the test data that one cannot identify with routine database queries.

2. DATA MINING TECHNIQUES

2.1 Introduction

Data Mining (also called Knowledge Discovery) is the process of analyzing data from different perspectives and summarizing it into useful information or relationships. It is also characterized as the process of finding correlations or patterns among dozens of fields in large relational databases. Data Mining as defined by Larose [1] *is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization to address the issue of information extraction from large data bases.*

Such extracted information is deemed both understandable and useful to the data owner. Organizations accumulate vast amounts of data in different formats and different databases, which are often located at more than one physical location. For Data Mining applications, only the elements that can be interpreted and converted into a computer usable form are treated as data. Data are defined as, *any facts, numbers, or text that can be processed by a computer*, to generate information. This may include:

a. Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
b. Non-operational data, such as industry sales, forecast data, and macro economic data
c. Meta data--data about the data itself, such as logical database design or data dictionary definitions

Further, the databases are combined, either by actual transfer of data or just logically with the use of database servers. From these, the data can be accessed on a real-time basis in ever increasing amounts and frequencies. Almost all of the medium to large organizations have built their own data banks (warehouses). The term data warehousing is defined in [1], *as a process of centralized data management and retrieval. It represents an ideal vision of maintaining a central repository of all organizational data.*

The organizations often rely on external agencies for proper use of the data. Thus, the data warehouses are subjected to use by various Data Warehousing agencies with the aim of retrieving useful bits of information. This information may not have any mathematical logical, or, for that matter, any sense at all, but may help the database researcher close in on the objective. Thus, information is defined by [1]; *are the patterns, associations, or relationships among all this data. Information can be converted into* ***knowledge*** *about historical patterns and future trends.*

Figure 2.1 shows the flow of data while being generated, stored and processed to get information which in turn becomes knowledge when interpreted.

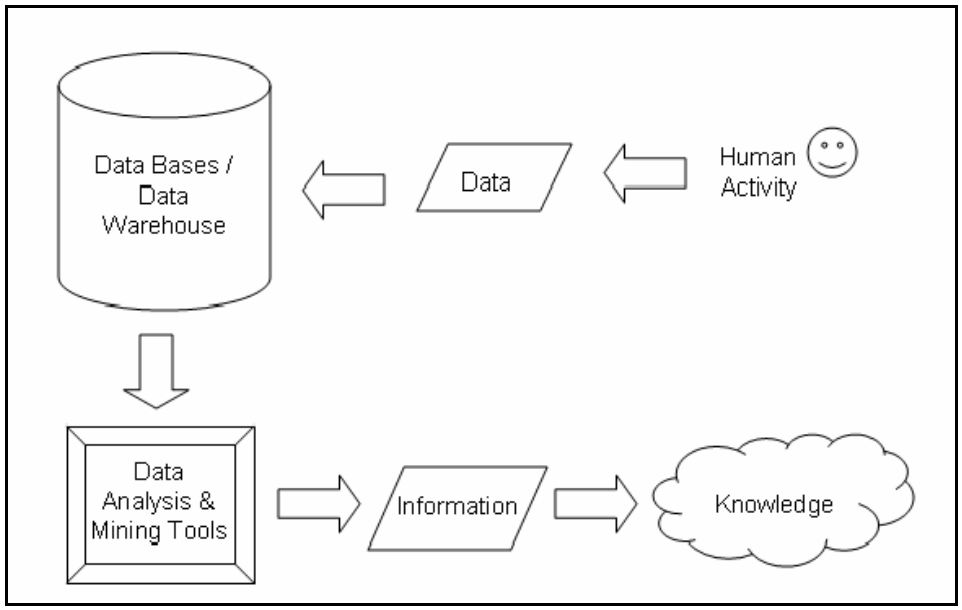

Figure 2.1: Flow of Data

2.2 Data Mining Methodology

Data Mining is typically carried out in six phases referred to as CRISP - DM (Cross-Industry Standard Process for Data Mining) [1].

1. Business Understanding: This is the initial phase of CRISP-DM. In this, one strives to understand the main objective of the Data Mining endeavor. A problem definition is established clearly stating its goals, restrictions and strategy.

2. Data Understanding: The primary focus during this phase is to evaluate data in terms of quality and quantity. Finally, an initial level classification is done on the database to get actionable data patterns.

3. Data Preparation: During this phase, these data are transformed into a suitable form. This often involves modifying/transforming the database as well as normalizing if needed.

4. Modeling: An appropriate model for the database is selected from the pool of options and optimized to suite the requirements. Often, more than one model is applied to meet the goals of the Data Mining project.

5. Evaluation: In this phase, the model is evaluated to check whether it meets all the required parameters, objectives, and constraints, and to monitor if any aspect is left unattended with respect to the pre-established goals.

6. Deployment: The model is implemented and detailed reports are generated based on the outcome. After brief analyses of the reports by the end user, possible ways of making more complex models are explored and to apply the model at more facilities.

Typical life cycle diagram of the CRISP_DM process is shown in the Figure 2.2 below.

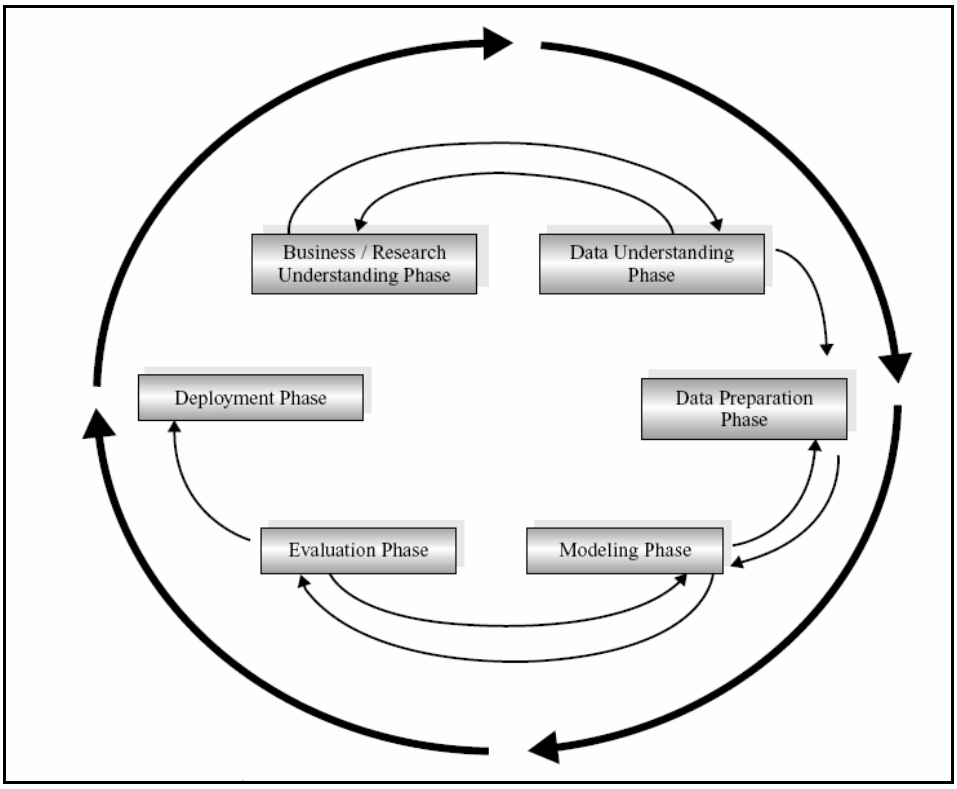

Figure 2.2: Cross-Industry Standard Process for Data Mining [1].

2.3 Data Mining Models

Selection of a Data Mining Technique is the most important aspect of any Data Mining endeavor. Based on the objective, one decides what data should be used from the pool of data, and how they should be processed. This is followed by analysis of that data in a predefined way, which paves the way for the selection of a particular algorithm or a set of algorithms. In many of the emerging applications, it is clear that no single approach is optimal and that multiple methods and approaches need to be used. Consequently, combining several modalities and classifiers or techniques is now a common practice [10].

The Data Mining models are categorized into different leaves. Further, each leaf signifies the relationship, if any, that is highlighted from the database. The techniques that

are actually applied may be a part of one or more leaves, and, based on which the categorization may vary. This sometimes is a cause of discord amongst the researchers. For most applications the Data Mining Models can be put into one of the six main categories: 1) Association, 2) Classification, 3) Clustering, 4) Prediction, 5) Sequence Discovery, and 6) Generalization. Figure 2.3 shows the Data Mining Models along with associated methods. The models are briefly discussed below:

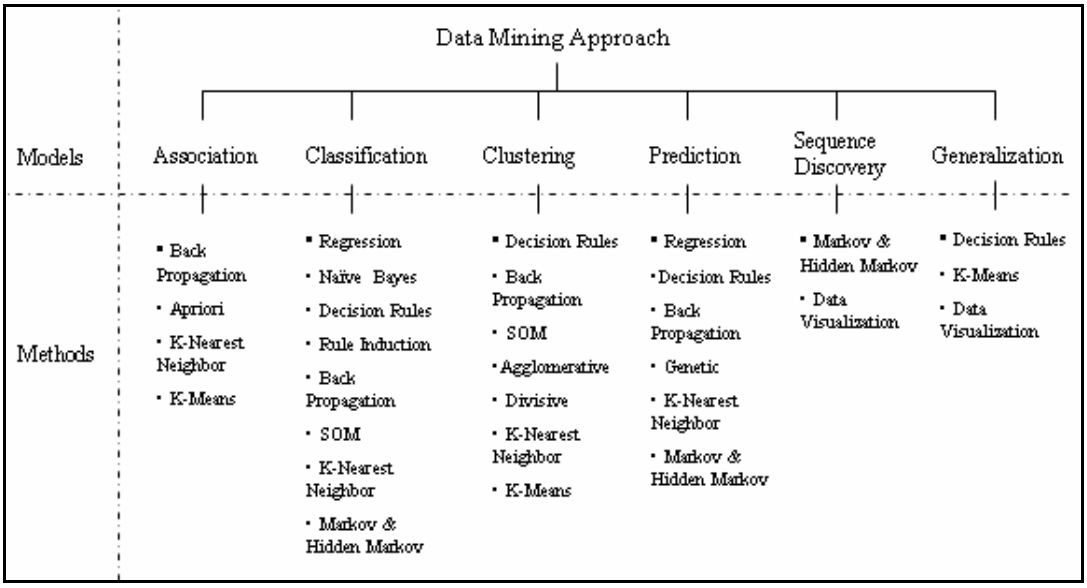

Figure 2.3: Data Mining Approach

1. Association:

Association rules depict the inter-relationship between the data items. These relationships often do not have any coherence with functional dependencies or correlation (Statistical Relationships). Interpretation of the outcome by this technique may be best achieved by behavioral-based reasoning rather than use of any engineering technique. The data are paired or transactional, often present in a date-time-based sequence. Partitioning of the data into training and validation sets purely depends on the type of algorithm being used for this technique. The sole aim is to come up with maximum number of item pairs with most number of transactions. The typical examples of association are the transaction data from a retail store, collected over the year. The interest is to find hidden relationships, if any, between any two items that are sold. For

example, the items can be formal trousers and a shoe polish, that otherwise do not have any statistical and/or functional relation, but makes sense if one follows the thinking of a person who might be buying the trousers for an interview and remembers to buy the shoe polish as well.

2. Classification:

Classification, as the name suggests, is the process of placing or assigning the categorical variables into predefined classes. An algorithm needs to be selected to place these data in categories. The decision rules are based on training data and then used to locate these data in pre-determined groups. The rules are further validated by the validation dataset. Classification falls under supervised or directed Data Mining technique. Most of the Data Mining techniques end up classifying these data in some way or other. The classification techniques can be best explained by an example. Let's say a restaurant chain could mine customer purchase data to determine when they visit and what they typically order. This information could be used to increase customer traffic by having daily specials. Also, the menu could be arranged to suit customer needs. Some of the issues in classification are:

a. Missing Data: Missing values in the database can have a significant impact on the decision rules. The missing data are filled by the mean value of the database or based on the business logic. This may lead to improper decision rules because they have tendency to pull the regression line towards themselves.

b. Measuring Accuracy: Accuracy is measured as the percentage of tupules that are placed into correct classes. This is often based on user intellect, the algorithms that are chosen and the quantity and quality of data. Data that overlap (belong to more than one class) can bear significant effect on the accuracy of the model.

c. Being a highly supervised technique user input is required to determine the number of levels, leaves, etc.

d. Finally, the size of training, validation, and, test data affect the outcome of the relationship significantly.

3. Clustering:

In clustering, the data items are grouped according to their logical relationships or natural groupings and a structure as a whole is generated. No clustering technique is universally applicable in uncovering the variety of structures present in multidimensional datasets; usually the 'best' applicable technique is selected [10]. There are no pre-defined groups, thus, clustering comes in the group of undirected Data Mining techniques. Each cluster is collection of homogeneous elements, which may be exclusive to that group, but are similar to each other. The presence of an element in a particular cluster may be definite or probabilistic. They might even have a hierarchical structure, having a crude division of elements at the highest level of the hierarchy, which is then refined to sub-clusters at lower levels. Moreover, each cluster may be different from other clusters. Clustering prior to application of other Data Mining techniques might reduce the complexity by dividing the space of elements [12]. These space partitions might exhibit improved results when mined separately. Clustering has found extensive application in the fields of marketing, Web mining, insurance, disaster planning, etc. Clustering techniques identify clusters, whereas the classification methods place data in pre-defined clusters. The main requirements of any clustering model are [3]:

a. It should be robust to deal with multiple attributes of data, have high dimensionality and scalability
b. It must successfully discover clusters even with arbitrary shape, but avoid reaching local minimum for a local function/decision
c. For some algorithms, the number of clusters or distance needs to be chosen scientifically
d. It should deal with noise and insensitivity better when compared to other models
e. For a given model, the user is responsible to check whether the clusters are real and interesting from an application point of view

4. Prediction:

In prediction, data are mined to anticipate behavior, patterns, and trends. This is often the outcome of the previous three basic models. The idea is that once the decision rules are generated through classification or clustering, those rules form the basis of the prediction model. Thus, all the techniques discussed earlier are capable of prediction to some extent. The error or the probability factor is considered while choosing any particular algorithm over another. A good example of prediction would be an outdoor equipment retailer predicting the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

5. Sequence Discovery:

Sequence discovery is used to determine sequential patterns in the data. These sequences are more often associations between various data fields, but they are essentially based on time and often follow a particular queue. This technique encompasses association rules as well as Markov concepts; hence not much can be elaborated on regarding this. For instance, if a person buys a CD player then he is bound to buy CDs for it sooner than later [2].

6. Generalization:

Generalization, also called Description or Summarization, pulls the data into subsets with their respective descriptions. Sometimes actual portions of the mined data are retrieved and based on that the subsets described above are created. Generalization is not a Data Mining method; it is the outcome of Data Mining technique.

Often algorithms are called by the name of the model they fit into or by the name of the algorithm. For example, an algorithm for finding clusters in the database is called a "Clustering Algorithm," whereas an algorithm used to establish association rules is called an "Apriori Algorithm."

2.4 Data Mining Methods

The common Data Mining methods are 1) Regression Analysis, 2) Naïve Bayes Classification, 3) Decision Rules, 4) Rule Induction, 5) Neural Networks, 6) Genetic Algorithm, 7) Association Rules, 8) Hierarchical Clustering, 9) Partitional Clustering, 10) Markov Model, and 11) Data Visualization. A brief description of each method is provided below:

1. Regression Analysis

Regression Analysis is used by the classification and prediction models. Statistical regression models are implemented that best fit the training data. The objective is to minimize the Sum of Squared Error (SSE). Once a model is developed, it is then validated with the validation dataset. The model is run on a new set of data (independent variables). Regression algorithms consider effects of interaction and correlation in data; they also measure trend in the data that many other Data Mining methods cannot. Even so, Logistic Regression has found more practical applications in many real-life scenarios [14]. It is a non-linear regression model where the response variables are discrete in nature, often qualitative for Data Mining, with a sigmoidal response function. Different response functions are used, namely, Probit, Logit, Log-Log, MLE, give the probability for the event. These functions often generate the odds or probability of the event under study.

2. Naïve Bayes Algorithm

The Naïve Bayes algorithm is based on the Bayesian theorem and is particularly well suited when the dimensionality of the inputs is high. Essentially it classifies a new record based on probabilities estimated from the training data. Despite its simplicity, Naïve Bayes can often outperform many sophisticated classification methods.

For example, a task could be to classify new cases as they arrive based on the existing types of objects with their respective probability of occurrence i.e. prior

probability. Then, for a new point in space, one can calculate the likelihood as, the number of objects in vicinity, divided by total objects of the same type. The final classification would then be produced by combining both sources of information, i.e., the prior and the likelihood, to form a probability of classification into a particular class (posterior probability) using the Bayes' rule. This algorithm has accuracy problems, based on the complexity of data, for continuous data.

3. Decision Rules

Decision Rules or decision trees are based on an inductive approach. Decision rules take an if/then approach and Decision trees develop a graph similar to a tree diagram. Decision trees are constructed in order to help in decisions that in turn generate rules for the classification of a new unclassified dataset. They also act as a tool for selection of variables for Neural Networks and Regression. Some of the methods are Iterative Dichotomizer 3 (ID3), Classification and Regression Trees (CART), and Chi Square Automatic Interaction Detection (CHAID). ID3 and CART segment the dataset by binary splits and CHAID does multi-way splits. Also, decision rule algorithms have difficulty with missing values (now imputed, i.e., removed or replaced), continuous data (now pruned), and most ignore correlations and interactions in data [2].

4. Rule Induction

Rule Induction algorithms work complementarily to the Decision Tree algorithm, concerning, extraction of useful rules from data. Each class is considered separately and then an attempt is made to cover all elements in that class. The Rule Induction is called the “Covering Approach,” because at each stage, rules would be determined that cover a set of the elements. Further, these algorithms test the rule being constructed, thus maximizing accuracy.

5. Neural Networks

Neural Networks algorithms originated from the research aimed at understanding biological information processing systems. Thus, the concept of Artificial Neural Networks (ANN) was born in an attempt to simulate the brain's processing system. Even though there is lack of understanding about how the brain works, the models that were built produced reasonably good results. Instead of using the central processor to analyze a number of instructions, the neural net software analyzes the data by passing them through several simulated processes that are interconnected.

A simple neural network consists of input, hidden, and output layers:

1. The Input Layer is comprised of the input nodes. The input data are scaled and then a model is built to calculate the scaled output values.

2. The Hidden Layer receives the data from the input nodes and transmits their combined value to the output node.

3. The Output Layer gets the inputs from the hidden layer and these inputs are used to calculate the output of the whole system. These outputs are scaled back to their original values.

The Neural Networks can be used for Data Mining applications [15]. Some of the popular Neural Network algorithms for Data Mining are: a) Back Propagation Algorithm that is effectively able to do all six main tasks of Data Mining, but primarily, used for prediction purposes, b) Kohonen, Self Organizing Maps is essentially a clustering method only, and has been extensively used for that.

Historically, Neural Networks are found to generate outcomes that are best amongst the group, but often by a small margin especially in case of Regression Algorithms and Decision Trees. At the same time, they use many more parameters as compared to any other method. Therefore, in practice there is a trade off with the use of Neural Networks and other Data Mining Algorithms. All Neural Networks are prone to

over-fitting the data; they require data in numerical or converted to numerical form. Moreover, their training is slow and, being "Black Box," relatively harder to interpret and use.

6. Genetic Algorithm

This optimization search technique is based on the concepts of natural evolution, sometimes considered as an automatically evolving special case of the Rule Induction algorithm [16]. Unlike other algorithms, the output is general yet specific to the case under study and is understandable mainly by the domain experts. Hence, it cannot be used to build generic applications for Data Mining. During each run, a large population of random chromosomes is created. When decoded, each one represents a different solution to the problem at hand. The following steps are repeated until a solution is found:

a. Test each chromosome with fitness function to see how good it is at solving the problem at hand and assign a fitness score accordingly

b. Select two members from the current population based on fitness. The Roulette wheel selection is a commonly used method where size of a section is proportional to fitness

c. Dependent on the crossover rate the bits from each chosen chromosome swap at a randomly chosen point

d. Step through the chosen chromosomes bits and flip depending on the mutation rates

e. Repeat step b, c, d until a new population of N members has been created

Thus a generalized solution to problem is reached. Genetic Algorithm is preferred for applications pertaining to relatively smaller databases.

7. Association Algorithm

Association algorithms generate candidate item sets from complete item sets; of those, the ones with infrequent sub-patterns are pruned. Thus, the generated candidate sets contain item sets that scan the whole transaction database to determine frequent item sets among the candidates. For determining frequent items in a fast manner, the algorithm stores candidate item sets. Unlike other algorithms, Association algorithms are designed to operate on databases containing transactions. The algorithms are generally simple to implement and reasonably efficient.

For example, the owner of a store that sells DVDs, VCD, books, and games may want to discover which of these items customers are likely to buy together. Customers in this particular store may like buying a DVD and a game in 10 out of every 100 transactions, or the sale of VCD may hardly ever be associated with a sale of a DVD. With the information above, the store could strive for more optimum placement of DVDs and games, as the sale of one of them may improve the chances of the sale of the other frequently associated item.

Some of the association methods are: Apriori, Partitioning Algorithm, Count Distribution Algorithm (CDA), and Data Distribution Algorithm (DDA) [2]. Of these, the Partitioning algorithms generate a large number of candidates for non-uniform data, thus increasing the computational complexity and resource consumption, but they are able to utilize the memory. The CDA and DDA often need parallel processing or even multi-processing for computation as they have large message traffic.

Advantages and Drawbacks

Advantages:

a. Decision Rule models are simple to understand and interpret (White Box).
b. They can handle both nominal and categorical data, once discrete before use in the model.
c. They do not need normalization and or removal of missing values.
d. They are robust and work well with large data.

Drawbacks:

a. Difficult to predict the value of the continuous attributes.
b. Reduced accuracy of classification for high dimensional data.
c. They are computationally expensive. Moreover, pruning algorithms add to the complexity as many candidate sub-trees are formed and compared.
d. Decision rules often lead to rectangular classification boxes that might not correspond well with the actual distribution of records.

8. Hierarchical Clustering

The two primary algorithms of this type are a) Agglomerative Algorithm and b) Divisive Algorithm [2].

Agglomerative Clustering

Agglomerative Clustering algorithms start with each individual item as its own cluster and then iteratively merge clusters until all items merge into one cluster as per the threshold distance criteria. The three sub-types of these algorithms are Single Link, Complete Link and Average Link algorithms.

Divisive Clustering

These algorithms behave in the way that is exactly opposite to the Agglomerative Clustering. Here, all the items are considered to be part of a single cluster and then split. All hierarchical clustering algorithms face excessive space and time constraints. Thus, they are difficult to use on large databases. Moreover, these algorithms cannot handle trends in data, whenever new data are added or old data removed, it has to be re-run.

9. Partitional Clustering

The two primary algorithms of this type of clustering are: a) K- Means Clustering, and b) Nearest Neighbor.

K – means Clustering

K-means Clustering are the most common unsupervised learning algorithms that are used for classification or partition of the data into clusters that are fixed a priori. The main idea is to pre-define centroids, one for each cluster, the artificial points in the spaces which represent a mean location of all the items of that particular cluster. These centroids should be placed in an intelligent way, as the algorithm is sensitive to initial selection, because their different location gives different results. So, the better choice is to place them as far as possible from each other. *Centroid is defined as the representative element of a cluster; it is the point whose parameter values are the average of the parameter values of every point in the current cluster.* The steps of the K-means algorithm are:

a. Randomly select some points to be the seeds for the centroids of the clusters.

b. Assign data items to the centroid closest to them, thus forming clusters of items.

c. Find new centroids for the items belonging to the same cluster.

d. If the centroids have changed their "coordinates," then start again from step b. Else, cluster detection is finished with all cluster memberships defined.

The K-means algorithm might not find the most optimal configuration, corresponding to the global objective function minimum. A simple approach is to compare the results of multiple runs with different seed selection and choose the best one as per given criterion, while taking care of overfitting. Also, the values in the dataset must all be numeric; categorical data must be transformed into numeric ones, and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space. Median values are often used to tackle outliers. Also, individual significance of the items in the group, towards the centroid, remains unknown.

Nearest Neighbor

These are memory-based algorithms, typically used for classification and prediction scenario. Each dataset consists of a set of independent values labeled by a set of dependent outcomes. They would be either continuous or categorical based on which regression or classification is done. For a new dependent value called query point, the outcome is estimated based on the KNN datasets. For regression, the predictions are based on averaging the outcomes of the k nearest neighbors and for classification; voting is used. Choice of "k" could be the most important factor of the algorithm which also acts as a smoothing parameter. A higher value for "k" increases the bias and reduces variance and vice versa also holds true.

The value of "k" is estimated by Cross-validation when model parameters are unknown. The data are divided into numbers of randomly drawn samples. For a fixed value of "k," the KNN model is used to make predictions and evaluate the error (SSE for regression or Accuracy for prediction). This is then successively applied to all possible sample choices. At the end, the errors are averaged to yield a measure of the stability of the model. The above steps are repeated for various "k" and the value with lowest error or highest classification accuracy is selected. It would be worth the effort, however only for fast queries and large numbers of items. Moreover, nearest neighbor search gets progressively harder as the dimensionality increases.

For prediction i.e. regression the typical score functions would be Euclidean, or other Weighted Distance measures. While for classification, a voting scheme is used. The KNN is often found to be vulnerable in case of equal instances and weights of variables.

10. Markov and Hidden Markov

Markov and Hidden Markov Algorithms are widely used to model sequential processes, and have achieved practical successes in many areas such as Web log mining, computational biology, speech recognition, robotics, fault diagnosis, and survival analysis [17]. A first-order Markov model contains a single variable, the state, and specifies the probability of each state and of transiting from one state to another. Hidden Markov

Models (HMMs) contain two variables: the hidden state and the observation. In addition to the transition probabilities, HMMs specify the probability of making each observation in each state. Because the number of parameters of a first-order Markov model is quadratic in the number of states (and higher for higher-order models), learning Markov models is feasible only in relatively small state spaces. This requirement makes them unsuitable for many Data Mining applications, which are concerned with very large state spaces. Still, their advantage is that they can be estimated statistically and adapted quickly. For example, they could be used to predict the probability of seeing a link on a Web page given a history of accessed links and also help generate a sequence of links that could be accessed by the user in the future.

11. Data Visualization

The visual interpretations of complex relationships in the multidimensional data are often termed as Data Visualization. Many new and innovative graphics tools are used to illustrate the data in space and, more importantly the data relationships. Data Visualization is mostly used in combination with other algorithms to make the results more presentable, readable, and understandable. Thus, it is an important part of Data Mining group, even though it does not analyze the data.

DMQL

With increased use of RDBMS (Relational Database Models) and recent advancements in SQL based querying tools, such as Oracle 9i and MS SQL Server 2005, more complicated tasks such as OLAP (Online Analytical Processing) and Data Mining Applications are being developed. A Data Mining Query Language (DMQL) has been proposed. This works only with RDBMS [2].

Table 2.1 classifies the various Data Mining algorithms according to problem type, namely, Association, Classification, Clustering, Prediction, Discovery, and Summarization.

Table 2.1: List of Data Mining Algorithms									
S. No.	**Methods**	**Problem Types**						**Input**	**Output**
		Association	Classification	Clustering	Prediction	Sequence Discovery	Description/ Summarization		
1	Regression Analysis		**Y**		**Y**			Numerical Values	Multivariate Equation
2	Naïve Bayes Classification Algorithm (Probabilistic)		**Y**					Numerical Values	Probability of outcomes belonging to particular classes.
3.1	Decision Rules or Trees (DT): ID3		**Y**	**Y**	**Y**		**Y**	Sample Computational Data, Numeric.	Decision Rules/Tree
3.2	DT: C4.5		**Y**		**Y**			Sample Computational Data,	Decision Rules/Tree
3.3	DT: Classification And Regression Trees (CART)		**Y**		**Y**			Sample Computational Data, Numeric.	Decision Rules/Binary Tree
3.4	DT: Chi Squared Automatic Interaction Detection (CHAID)		**Y**		**Y**			Sample Computational Data, Numeric.	Decision Rules/Tree
3.5	DT: Quick Unbiased Efficient Statistical Tree (QUEST)		**Y**		**Y**			Sample Computational Data, Numeric.	Decision Rules/Tree
4	Rule Induction		**Y**		**Y**			Sample Computational Data, Numeric.	Decision Rules/Tree
5.1	NN: Back Propagation Algorithm	**Y**	**Y**	**Y**	**Y**			Feature of pattern.	Classification, Prediction, Clustering and Association results.

S. No.	Methods	Problem Types						Input	Output
		Association	Classification	Clustering	Prediction	Sequence Discovery	Description/ Summarization		
5.2	NN: Self Organizing Map - Kohonen		**Y**	**Y**				Feature of pattern.	Clusters of data.
6	Genetic Algorithm				**Y**			Binary streams.	Set of rules.
7	Association Algorithm (Apriori)	**Y**				**Y**		Transactional Data	Freq. of Associations of features.
8.1	Agglomerative Algorithm (Hierarchical Clustering)			**Y**				Sample feature data.	Clusters of data.
8.2	Divisive Algorithm (Hierarchical Clustering)			**Y**				Sample feature data.	Clusters of data.
9.1	k - Nearest Neighbor (Partitional Clustering)	**Y**	**Y**	**Y**	**Y**			Features of the predictor and/or classifier variables.	Classification or grouping of datasets based on weighted distance from the target or query (centroid).
9.2	k - Means Clustering (Partitional Clustering)	**Y**		**Y**			**Y**	Features of the predictor and/or classifier variables.	Clusters or grouping of datasets based on similarities or closest to centroid.
10	Markov and Hidden Markov Model		**Y**		**Y**	**Y**		Instance based data.	Sequence or probability of a sequence.
11	Data Visualization					**Y**	**Y**	Not Considered	Not Considered

2.5 Algorithm Structure

Section 2.3 described how the Data Mining models are categorized. This section describes the different components of Data Mining algorithms. The four components of a typical Data Mining algorithm are [13]:

1. Pattern Structure:

Pattern structure is the underlying structure or functional form that is part of the algorithm. Based on the score or selection criterion, data similar to the pattern structures are sought from the data. They are often called the relevant condition, constraint or parameter that will be computed by the algorithm. The name "pattern structure" is sometimes replaced by the term "model structure."

2. Score Function:

This quantifies how well a given pattern or parameter fits the given dataset. Thus, based on the error that is generated, modifications are made to the algorithm. Further, the range of the input values to the model, in the form of constraints of the function, can be determined. They often serve the purpose of comparing the utility or efficiency of one method over the other, thus helping to choose one model over another. Finally, it is desirable that the score function should be robust and not be sensitive to the input values. There are many types of score functions, including Squared Error, Gini Coefficient, Euclidean Distance, and Entropy.

3. Optimization and Search Function:

The task of the optimization function is to determine the "best" set of values from the given data depending on the objective. The best set essentially means the values that minimize or maximize the score function as desired by the Data Mining logic. The task of finding an interesting pattern is essentially the task of a search function. The search function may make use of searching techniques such as binary search, index search and more importantly using different heuristic search techniques.

4. Data Management Strategy

This concerns efficient handling of data during the optimization and search operation. Until recently, this had little bearing on the Data Mining model, but now significant emphasis is put on the data that are being managed in order to improve the overall efficiency of the Data Mining algorithm. Moreover, the way (format, order, etc.) in which data are to be arranged from the user or algorithm viewpoint is also important. This is supported by fact that with increase in the size of data, it takes more and more time to train, validate and actually mine them.

2.6 Summary

The significance of the Data Mining field can be attributed to:

a. Explosive growth in data.
b. Advancement in data gathering mechanisms. Superior Internet and intranet facilities.
c. Proportionate increase in data storage and computation capability.
d. Competitive pressure to maintain market share in this era of globalization.
e. Advancement in health care facilities.
f. Security needs of the community, country and the world.

Some issues concerning Data Mining are:

a. It is an open-ended search which does not guaranty a solution to the problem.
b. Most algorithms require significantly skilled human input.
c. The outcome of the Data Mining procedure is highly dependent on the data quality and quantity for the given objective.
d. The significance of the outcome has to be gauged by the end user, while keeping the objective in mind.
e. The returns of Data Mining in terms of cost, time, and effort are a matter of dispute.

3. DATABASE DESIGN AND IMPLEMENTATION

3.1 Introduction

For the crime database, various data fields were identified to cover real world scenarios. The inputs were taken from the Forensic Information Management System (FIMS) that is being developed for West Virginia State Police Department [21] and Web sources of various crime investigation agencies [22]. The entities that were identified are, 1) Offense, 2) Suspect, 3) Victim, 4) Convict, 5) Investigating Agency, 6) Officer In-charge, 7) Charges (against convict), and 8) Verdict. Four groups of tables as shown in Table 3.1 were created (Data Tables, Code Tables, Link Tables, and Query Tables). The various tables under each group are shown below.

Table 3.1: Tables in Crime Database

S. No.	Data Tables	S. No.	Code Tables	S. No.	Link Tables	S. No	Query Tables
1	dtOffense	1	ctOffenseType	1	ltLnkOffSus	1	tblQuery
2	dtSuspect_Physical	2	ctDrugListType	2	ltLnkOffVic		
3	dtSuspect_Alias	3	ctAgencyType	3	ltLnkCon		
4	dtSuspect_Variable	4	ctOfficerTitle				
5	dtVictim_Physical	5	ctCountyType				
6	dtVictim_Alias	6	ctStateType				
7	dtVictim_Variable	7	ctCountryType				
8	dtAgencyInfo	8	ctRaceType				
9	dtOfficerInfo	9	ctGenderType				
10	dtOfficerWorkInfo	10	ctEyeColorType				
11	dtConvict	11	ctHairColorType				
12	dtCharges	12	ctFinCondtnType				
13	dtVerdict	13	ctBehvAspctType				

3.2 Data Tables

The fields of the data tables were built from various sources as stated earlier. The tables are populated as the cases are registered by the crime investigating agency. The various fields of the "dtOffense" table are shown in Table 3.2. This table has details related to offense such as registration, time, type and address of offense as well as brief description of offense.

Table 3.2: Fields of "dtOffense" Table

Field Name	Data Type	Length	Description
ACN	nvarchar	50	Agency Case Number
AgencyID	nvarchar	50	Agency Identification Number
OfficerID	nvarchar	50	Officer's Identification Number
EntryDate	datetime	8	Entry date of case
ModifDate	datetime	8	Modification date
OffenseDate	datetime	8	Offense date
OffenseTime	datetime	8	Offense time
OffenseDay	nvarchar	50	Offense day
OffenseType	nvarchar	50	Offense type
Street	nvarchar	50	Address of offense
City	nvarchar	50	City of Offense
County	nvarchar	50	County of Offense
State	nvarchar	50	State of Offense
Country	nvarchar	50	Country of Offense
ZipCode	nvarchar	50	Zip Code of Offense
OffenseDescp	nvarchar	500	Brief description of offense
CaseStatus	nvarchar	50	Status of the case

The fields of the "dtSuspect_Physical" table are shown in Table 3.3. This table stores the physical attributes (non-changeable) of a suspect in the database and a suspect identification number is generated.

Table 3.3: Fields of "dtSuspect_Physical" Table

Field Name	Data Type	Length	Description
SuspectID	nvarchar	50	Identification Number of suspect
EntryDate	datetime	8	Entry date of record
ModifDate	datetime	8	Modification date of record
SSN	nvarchar	50	Social security of the suspect
LName	nvarchar	50	Last Name of the suspect
FName	nvarchar	50	First Name of the suspect
MName	nvarchar	50	Middle Name of the suspect
Suffix	nvarchar	50	Suffix of the suspect
DOB	nvarchar	8	Date of birth
POB	nvarchar	50	Place of birth
Age	nvarchar	50	Age of the suspect
Height	nvarchar	50	Height of the suspect
Weight	nvarchar	50	Weight of the suspect
Dexterity	nvarchar	50	Right/Left handedness of the suspect
EyeColor	nvarchar	50	Eye Color of the suspect
HairColor	nvarchar	50	Hair Color of the suspect
SkinColor	nvarchar	50	Skin Color of the suspect
FingerClass	nvarchar	50	Finger Print Classification of the suspect
FingerPatt	nvarchar	50	Finger Print Pattern of the suspect
IndvMarks	nvarchar	500	Marks on individual's body of the suspect
Race	nvarchar	50	Race of the suspect
Gender	nvarchar	50	Gender of the suspect

The fields of the table "dtSuspect_Alias" are shown in Table 3.4. This table stores various aliases of the suspect in the database for the same suspect identification number.

Table 3.4: Fields of "dtSuspect_Alias" Table

Field Name	Data Type	Length	Description
SuspectID	nvarchar	50	Identification Number of suspect.
AliasNum	nvarchar	50	Alias Number of suspect
LastName	nvarchar	50	Last Name of the suspect
FirstName	nvarchar	50	First Name of the suspect
MiddleName	nvarchar	50	Middle Name of the suspect
Suffix	nvarchar	50	Suffix of the suspect

The fields of the "dtSuspect_Variable" table are shown in Table 3.5. This table stores variable characteristics (changeable) of a suspect in the database based on suspect identification number and alias number.

Table 3.5: Fields of "dtSuspect_Variable" Table

Field Name	Data Type	Length	Description
SuspectID	nvarchar	50	Identification Number of Suspect
AliasNum	nvarchar	50	Alias Number of Suspect
EntryLevel	nvarchar	50	Level of record entry for given alias
EntryDate	datetime	8	Date of record entry
PassportNo	nvarchar	50	Passport No of suspect
Citizenship	nvarchar	50	Citizenship of suspect
LiscenceNo	nvarchar	50	License No of suspect
Religion	nvarchar	50	Religion of suspect
Occupation	nvarchar	50	Occupation of suspect
Street	nvarchar	50	Street of suspect
City	nvarchar	50	City of suspect
County	nvarchar	50	County of suspect
State	nvarchar	50	State of suspect
Country	nvarchar	50	Country of suspect
ZipCode	nvarchar	50	Zip Code of suspect
Phone	nvarchar	50	Phone No of suspect
EmerContact	nvarchar	50	Emergency contact of suspect
FinCondtn	nvarchar	50	Financial condition of suspect
FinDescp	nvarchar	50	Financial description of suspect
BehvAspect	nvarchar	50	Behavioral aspect of suspect
BehvDescp	nvarchar	50	Behavioral description of suspect
DefnTrain	nvarchar	50	Defense training if any
WeapTrain	nvarchar	50	Weapons training if any
DrugHist	nvarchar	50	Drug history if any
DrugType	nvarchar	50	Drug type associated with if any
Charge1	nvarchar	50	Charge1 if any
Charge2	nvarchar	50	Charge2 if any
Charge3	nvarchar	50	Charge3 if any
Status	nvarchar	50	Status as in prior conviction etc

The fields of the "dtVictim_Physical" table are shown in Table 3.6. This table stores the physical attributes (non-changeable) of a victim in the database and a victim identification number is generated.

Table 3.6: Fields of "dtVictim_Physical" Table

Field Name	Data Type	Length	Description
VictimID	nvarchar	50	Identification Number of victim
EntryDate	datetime	8	Entry date of record
ModifDate	datetime	8	Modification date of record
SSN	nvarchar	50	Social security of the victim
LName	nvarchar	50	Last Name of the victim
FName	nvarchar	50	First Name of the victim
MName	nvarchar	50	Middle Name of the victim
Suffix	nvarchar	50	Suffix of the victim
DOB	nvarchar	8	Date of birth
POB	nvarchar	50	Place of birth
Age	nvarchar	50	Age of the victim
Height	nvarchar	50	Height of the victim
Weight	nvarchar	50	Weight of the victim
Dexterity	nvarchar	50	Right/Left handedness of the victim
EyeColor	nvarchar	50	Eye Color of the victim
HairColor	nvarchar	50	Hair Color of the victim
SkinColor	nvarchar	50	Skin Color of the victim
FingerClass	nvarchar	50	Finger Print Classification of the victim
FingerPatt	nvarchar	50	Finger Print Pattern of the victim
IndvMarks	nvarchar	500	Marks on individual's body of the victim
Race	nvarchar	50	Race of the victim
Gender	nvarchar	50	Gender of the victim

The fields of the "dtVictim_Alias" table are shown in Table 3.7. This table stores the aliases of the victim in the database for the same victim identification number.

Table 3.7: Fields of "dtVictim_Alias" Table

Field Name	Data Type	Length	Description
VictimID	nvarchar	50	Identification Number of victim
AliasNum	nvarchar	50	Alias Number of victim
LastName	nvarchar	50	Last Name of the victim
FirstName	nvarchar	50	First Name of the victim
MiddleName	nvarchar	50	Middle Name of the victim
Suffix	nvarchar	50	Suffix of the victim

The fields of the "dtVictim_Variable" table are shown in Table 3.8. This table stores variable characteristics (changeable) of a victim in the database based on identification number and alias number.

Table 3.8: Fields of "dtVictim_Variable" Table

Field Name	Data Type	Length	Description
Victim ID	nvarchar	50	Identification Number of victim
AliasNum	nvarchar	50	Alias Number of victim
EntryLevel	nvarchar	50	Level of record entry for given alias
EntryDate	datetime	8	Date of record entry
PassportNo	nvarchar	50	Passport No of victim
Citizenship	nvarchar	50	Citizenship of victim
LiscenceNo	nvarchar	50	License No of victim
Religion	nvarchar	50	Religion of victim
Occupation	nvarchar	50	Occupation of victim
Street	nvarchar	50	Street of victim
City	nvarchar	50	City of victim
County	nvarchar	50	County of victim
State	nvarchar	50	State of victim
Country	nvarchar	50	Country of victim
ZipCode	nvarchar	50	Zip Code of victim
Phone	nvarchar	50	Phone No of victim
EmerContact	nvarchar	50	Emergency contact of victim
FinCondtn	nvarchar	50	Financial condition of victim
FinDescp	nvarchar	50	Financial description of victim
BehvAspect	nvarchar	50	Behavioral aspect of victim
BehvDescp	nvarchar	50	Behavioral description of victim
DefnTrain	nvarchar	50	Defense training if any
WeapTrain	nvarchar	50	Weapons training if any
DrugHist	nvarchar	50	Drug history if any
DrugType	nvarchar	50	Drug type associated with if any
Charge1	nvarchar	50	Charge1 if any
Charge2	nvarchar	50	Charge2 if any
Charge3	nvarchar	50	Charge3 if any
Status	nvarchar	50	Status as in prior conviction etc

The fields of the "dtAgencyInfo" table are shown in Table 3.9. This table stores information pertaining to the different agencies involved in solving the crime.

Table 3.9: Fields of "dtAgencyInfo" Table

Field Name	Data Type	Length	Description
AgencyID	nvarchar	50	Agency Identification Number
AgencyName	nvarchar	50	Agency Name
AgencyType	nvarchar	50	Type of agency Ex. Forensic, Investigation etc.
Street	nvarchar	50	Street of the agency
City	nvarchar	50	City of the agency
County	nvarchar	50	County of the agency
State	nvarchar	50	State of the agency
Country	nvarchar	50	Country of the agency
ZipCode	nvarchar	50	Zip Code of the agency
Phone	nvarchar	50	Phone number of the agency
Contact	nvarchar	50	Contact person at the agency
Notes	nvarchar	500	Important notes if any

The fields of the "dtOfficerInfo" table are shown in Table 3.10. This table stores the personal information about the crime investigating and other security officers.

Table 3.10: Fields of "dtOfficerInfo" Table

Field Name	Data Type	Length	Description
OfficerID	nvarchar	50	Identification Number of the officer
LName	nvarchar	50	Last Name of the officer
FName	nvarchar	50	First Name of the officer
MName	nvarchar	50	Middle Name of the officer
Suffix	nvarchar	50	Suffix of the officer
SSN	nvarchar	50	Social security number of the officer
DOB	datetime	8	Date of birth of the officer
Gender	nvarchar	50	Gender of the officer
Race	nvarchar	50	Race of the officer
Street	nvarchar	50	Street of the officer
City	nvarchar	50	City of the officer
County	nvarchar	50	County of the officer
State	nvarchar	50	State of the officer
Country	nvarchar	50	Country of the officer
ZipCode	nvarchar	50	Zip Code of the officer
Phone	nvarchar	50	Phone number of the officer
Notes	nvarchar	500	Important notes if any

The fields of the "dtOfficerWorkInfo" table are shown in Table 3.11. This table stores the professional information about the crime investigating and other security officers.

Table 3.11: Fields of "dtOfficerWorkInfo" Table

Field Name	Data Type	Length	Description
OfficerID	nvarchar	50	Identification Number of the officer
EntryDate	datetime	8	Date of record entry
ModifDate	datetime	8	Date of record modification
JobTitle	nvarchar	50	Title of the officer
Dept	nvarchar	50	Department of the officer
HireDate	datetime	8	Date of hire of the officer
TitleDate	datetime	8	Date of current title
Status	nvarchar	50	Status of the officer
Specialization	nvarchar	500	Specialization of the Officer
Performance	nvarchar	500	Performance of the Officer

The fields of the "dtConvict" table are shown in Table 3.12. This table stores the official information about the individuals convicted of crime.

Table 3.12: Fields of "dtConvict" Table

Field Name	Data Type	Length	Description
ConvictID	nvarchar	50	Identification Number of the convict
ArrestDate	datetime	8	Arrest date of the convict
CourtAssigned	nvarchar	50	Court assigned for trial
Trial	nvarchar	50	Type of trial
Counsel	nvarchar	50	Counsel status, type and name
PurposeCode	nvarchar	50	Purpose Code of the convict
OBTS	nvarchar	50	OBTS of the convict
SuspectID	nvarchar	50	Suspect Identification if matches
VictimID	nvarchar	50	Victim Identification if matches

The fields of the "dtConvict" table are shown in Table 3.13. This table stores the charges and related information of the individual convicted of crime.

Table 3.13: Fields of "dtCharges" Table

Field Name	Data Type	Length	Description
ConvictID	nvarchar	50	Convict's Identification Number
ChargeNum	nvarchar	50	Charge Number
Charge	nvarchar	50	Type of charge
Motivation	nvarchar	50	Motivation for crime
ProsecData	nvarchar	50	Data available with prosecution
CourtData	nvarchar	50	Data available with court
ConvictPlea	nvarchar	50	Convict's plea
Status	nvarchar	50	Guilty / Not Guilty

The fields of the "dtVerdict" table are shown in Table 3.14. This table stores the verdict against individual charge of the individual convicted of crime.

Table 3.14: Fields of "dtVerdict" Table

Field Name	Data Type	Length	Description
ConvictID	nvarchar	50	Convict Identification Number
ChargeNum	nvarchar	50	Charge Number
Statute	nvarchar	50	Statute
StatueDate	datetime	8	Date of statute
StatuteLevel	nvarchar	50	Level or Degree of statute
StatuteDescp	nvarchar	50	Description of Statute
Sentence	nvarchar	50	Type of sentence
SentenceDate	datetime	8	Date of sentence
SentencePlace	nvarchar	50	Place of sentence
FinalDispo	nvarchar	50	Final Disposition
DispoDate	datetime	8	Date of disposition
MdtRelDate	datetime	8	Mandatory release date
MdtExpDate	datetime	8	Mandatory expiry date
SupRelDate	datetime	8	Supplementary release date
SupExpDate	datetime	8	Supplementary expiry date
SupRelTerms	nvarchar	50	Supplementary release terms
SupStatus	nvarchar	50	Supplementary status
ParoleDate	datetime	8	Parole date
ParoleExpDate	datetime	8	Parole expiration date
ProbDate	datetime	8	Probation date

ProbExpDate	datetime	8	Probation expiration date
SPTDate	datetime	8	SPT date
SPTExpDate	datetime	8	SPT expiration date
PTDDate	datetime	8	PTD date
PTDExpDate	datetime	8	PTD expiration date
StatusWanted	nvarchar	50	Convict status

3.3 Code Tables

These tables are used to populate the fields of the data tables. They are also termed as "Filler Tables" as they do not serve any direct purpose other than the aforesaid task. The fields of the "ctAgencyType" table are shown in Table 3.15. This table stores different agency types.

Table 3.15: Fields of "ctAgencyType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
AgencyType	nvarchar	50	Types of Agency

The fields of the "ctOfficerType" table are shown in Table 3.16. This table stores different officer designations.

Table 3.16: Fields of "ctOfficerType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
OfficerTitle	nvarchar	50	Types of Officer Title

The fields of the "ctCountyType" table are shown in Table 3.16. This table stores all names of the counties in West Virginia.

Table 3.17: Fields of "ctCountyType" Table			
Field Name	**Data Type**	**Length**	**Description**
ID	nvarchar	50	Serial Number
CountyType	nvarchar	50	Type of Counties

The fields of the "ctStateType" table are shown in Table 3.18. This table stores names of all states in US. For now, this has only single id for the State of West Virginia.

Table 3.18: Fields of "ctStateType" Table			
Field Name	**Data Type**	**Length**	**Description**
ID	nvarchar	50	Serial Number
StateName	nvarchar	50	Type of States

The fields of the "ctCountryType" table are shown in Table 3.19. This table stores names of all the countries in the world. For now, this has only single id for the US.

Table 3.19: Fields of "ctCountryType" Table			
Field Name	**Data Type**	**Length**	**Description**
ID	nvarchar	50	Serial Number
CountryName	nvarchar	50	Type of Countries

The fields of the "ctRaceType" table are shown in Table 3.20. This table stores different races as identified by the FBI.

Table 3.20: Fields of "ctRaceType" Table			
Field Name	**Data Type**	**Length**	**Description**
ID	nvarchar	50	Serial Number
RaceType	nvarchar	50	Type of Race

The fields of the "ctGenderType" table are shown in Table 3.21. This table stores different genders as identified by the FBI.

Table 3.21: Fields of "ctGenderType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
GenderType	nvarchar	50	Type of Gender

The fields of the "ctCourtType" table are shown in Table 3.22. This table stores names of different courts in US.

Table 3.22: Fields of "ctCourtType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
CourtType	nvarchar	50	Type of Court

The fields of the "ctEyeColorType" table are shown in Table 3.23. This table stores eye colors as identified by the FBI.

Table 3.23: Fields of "ctEyeColorType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
EyeType	nvarchar	50	Eye Color Type

The fields of the "ctHairType" table are shown in Table 3.24. This table stores hair colors as identified by the FBI.

Table 3.24: Fields of "ctHairType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
HairType	nvarchar	50	Hair Color Type

The fields of the "ctFinCondtnType" table are shown in Table 3.25. This table stores attributes related to the financial condition of the individual.

Table 3.25: Fields of "ctFinCondtnType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
FinCondtnType	nvarchar	50	Types of Financial Condition

The fields of the "ctBehvAspectType" table are shown in Table 3.26. This table stores attributes related to the behavioral aspects of the individual.

Table 3.26: Fields of "ctBehvAspectType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
BehvAspectType	nvarchar	50	Types of Behavioral Aspect

The fields of the "ctOffenseType" table are shown in Table 3.27. This table stores all the offense types as identified by different investigation agencies.

Table 3.27: Fields of "ctOffenseType" Table

Field Name	Data Type	Length	Description
ID	nvarchar	50	Serial Number
OffenseType	nvarchar	50	Types of Offense

3.4 Link Tables

The link or linker tables are used to establish many-to-many (logical) relationship between the data tables. The individual tables make one-to-many relation with these tables, and while doing so maintain a many-to-many relation with one another. The linker tables are described below. The fields of the "ltLnkOffSus" table are shown in Table 3.28. This table acts as a link between the table "dtOffense" and "dtSuspect_Physical" table.

Table 3.28: Fields of "ltLnkOffSus" Table

Field Name	Data Type	Length	Description
ACN	nvarchar	50	Agency Case Number
SuspectID	nvarchar	50	Suspect Identification Number

The fields of the "ltLnkOffVic" table are shown in Table 3.29. This table acts as a link between the table "dtOffense" and "dtVictim_Physical" table.

Table 3.29: Fields of "ltLnkOffVic" Table

Field Name	Data Type	Length	Description
ACN	nvarchar	50	Agency Case Number
VictimID	nvarchar	50	Victim Identification Number

The fields of the "ltLnkOffVic" table are shown in Table 3.30. This table acts as a link between the table "dtOffense" and "dtConvict" table.

Table 3.30: Fields of "ltLnkOffCon" Table

Field Name	Data Type	Length	Description
ACN	nvarchar	50	Agency Case Number
ConvictID	nvarchar	50	Convict Identification Number

3.5 Query Table

Form the above tables, additional table can be created for the "Knowledge Discovery" purpose, called "tblQuery". This table will be built by extracting important fields from various tables of the database. Initially, the table will be populated with the data specifically prepared for it. In a production system this table would be populated simultaneously with the actual tables of the database. The focus of the initial table is to find typical crime patterns that were carried out within the state boundaries. The fields of the "tblQuery" table are shown in Table 3.31.

Table 3.31: Fields of "tblQuery" Table

Field Name	Data Type	Length	Description
ACN	nvarchar	50	Agency Case Number
OffYear	datetime	8	Offense Year
OffMonth	nvarchar	50	Offense Month
OffDate	nvarchar	50	Offense Date
OffHour	nvarchar	50	Offense Hour
OffMinutes	nvarchar	50	Offense Minutes
OffDay	nvarchar	50	Offense Day
County	nvarchar	50	Offense County
OffType	nvarchar	50	Offense Type
Gender	nvarchar	50	Convict's Gender
Race	nvarchar	50	Convict's Race
Age	nvarchar	50	Convict's Age
CriminalBkgd	nvarchar	50	Convict's Previous Crime Status (Yes / No)

The salient features of the Query Table are:

a. Important fields, based on domain knowledge and inputs from security agencies, were duplicated from the database into the "Query Table".

b. These fields were essentially put into single table to ensure the integrity and normalization constraints.

c. While doing so, the privacy and security of the victim, suspect and the convicts are ensured.

d. All fields of the table are of type “nominal” and “non-ordinal”; also most were “Null” able. Moreover, there are no “Dependent – Independent” variable relationship in the data.

e. The data entries are sequential in terms of date of registration and not as per actual event.

f. Based on the conditions Query table would be populated first to have random data in space. Then a pattern would be fed to it. Finally, all the tools would be used to find that pattern and the outcome would be reported.

3.6 Assumptions

The data tables had predefined relations between them, and to express them, certain database constraints are required [18]. These constraints provide data stability and at the same time cover security issues. The various assumptions were:

a. One criminal offense can be registered in many agencies.

b. One criminal offense is assigned to one investigating officer only. However, one investigating officer can handle multiple criminal offenses.

c. One criminal offense can include multiple suspects, victims and convicts. They may have multiple aliases having additional set of information.

d. One convict can face multiple criminal charges and each criminal charge will have a single verdict.

3.7 Database

The crime database was created from various tables described below. The relation between different data entities can be "one-to-one" or "one-to-many". The "key" symbol represents primary key relationship and the "∞" symbol represents one-to-many relationship. The relationships between various tables of the database are shown in the Figure 3.1 as below.

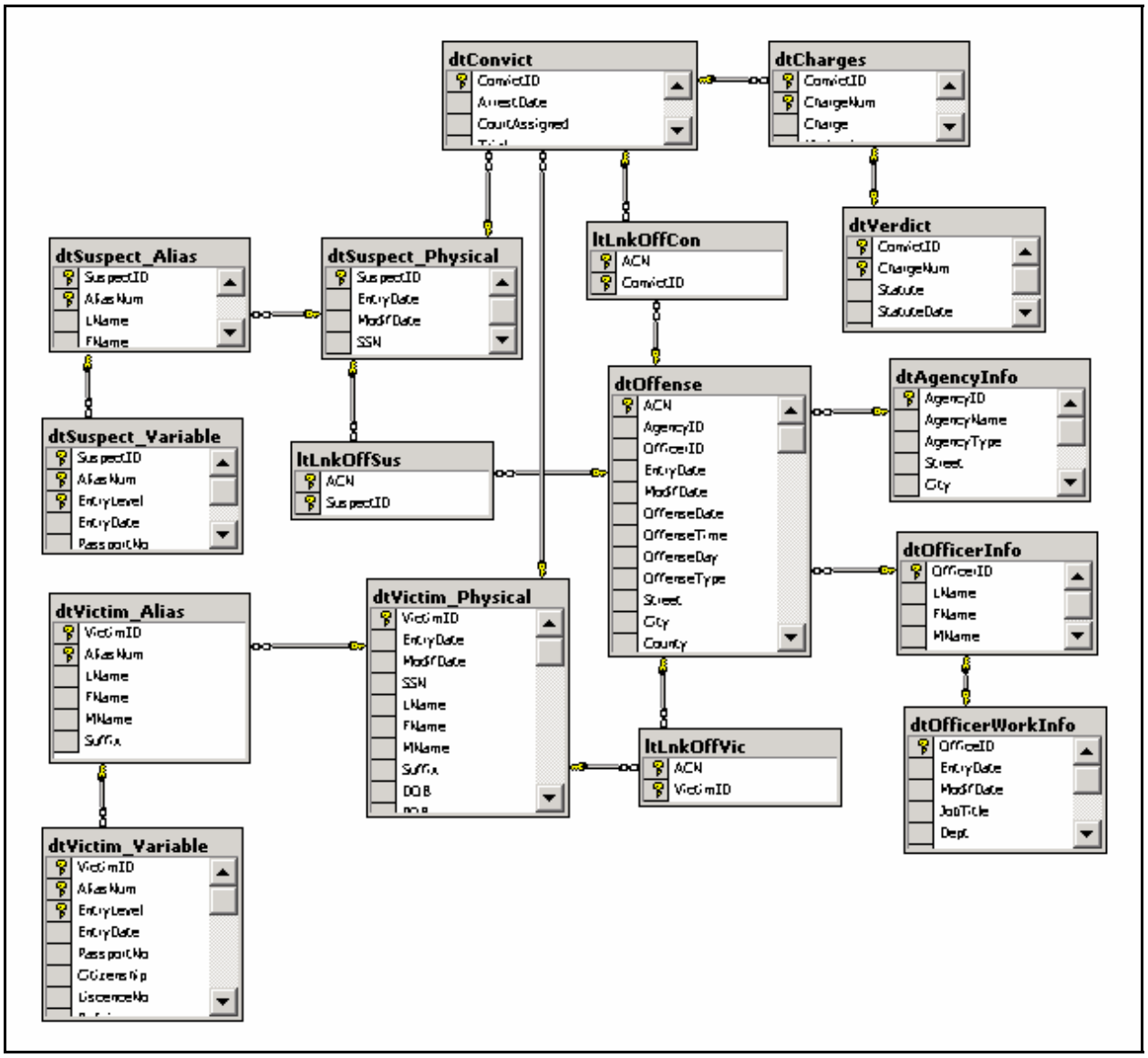

Figure 3.1: Database Diagram

4. USER INTERFACE DESIGN AND IMPLEMENTATION

4.1 Introduction

This chapter describes the design and implementation of the user interface. The task was to create Data Mining techniques on the query table that was extracted from the crime database. The table can be searched with tools such as the Search Engine, Expert Search or with the Data Mining techniques. The forms (front end) were designed in ASP.NET, thus giving the user Web-based applicability. Moreover, a Server (Windows) based application were implemented that enabled user to use WEKA for visualization of data.

4.2 Query Design

Querying application was designed in "Structured Query Language" and was primarily of "Select" type. The outcome of query was displayed with the help of "Data Grid". The user had a choice of selecting:

a. One or all fields along with their specific value.

b. User could assign conditions such as AND, OR, NOT and its combinations to each of the fields.

Based on user input a run time query is generated and that will be sent through the "Data Adapter" to the database and the outcome of the query displayed.

4.3 Query Implementation

The user can choose different fields, with condition and value selection. Figure 4.1 shows the user selected "Year" as 2005 and on pressing the "Search" button, the generated SQL query as well as the outcome are displayed.

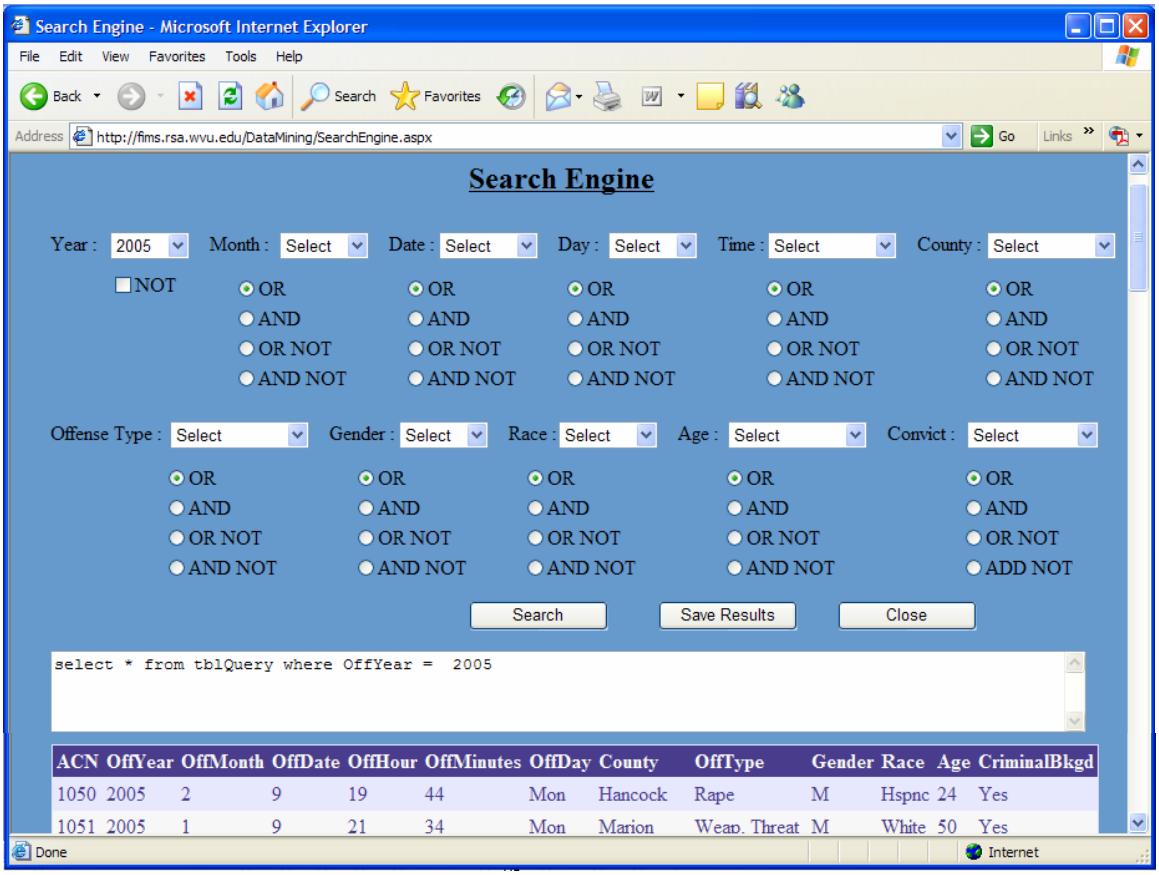

Figure 4.1: Search Engine

The Search Engine was able to generate basic / general queries and displayed the results that were queried by the user. But, it was unable to search multiple and simultaneous "If" statements (queries). Also, it was unable to offer multiple selections on one (same) field. These issues were primarily due to the user interface and data display constraints.

For that, a better user interface was designed and implemented. With this, the user had the flexibility to query the database with complex queries and also view only the desired information.

4.4 Modified Query Implementation

The modified query scheme, "Expert Search" is shown in Figure 4.2.

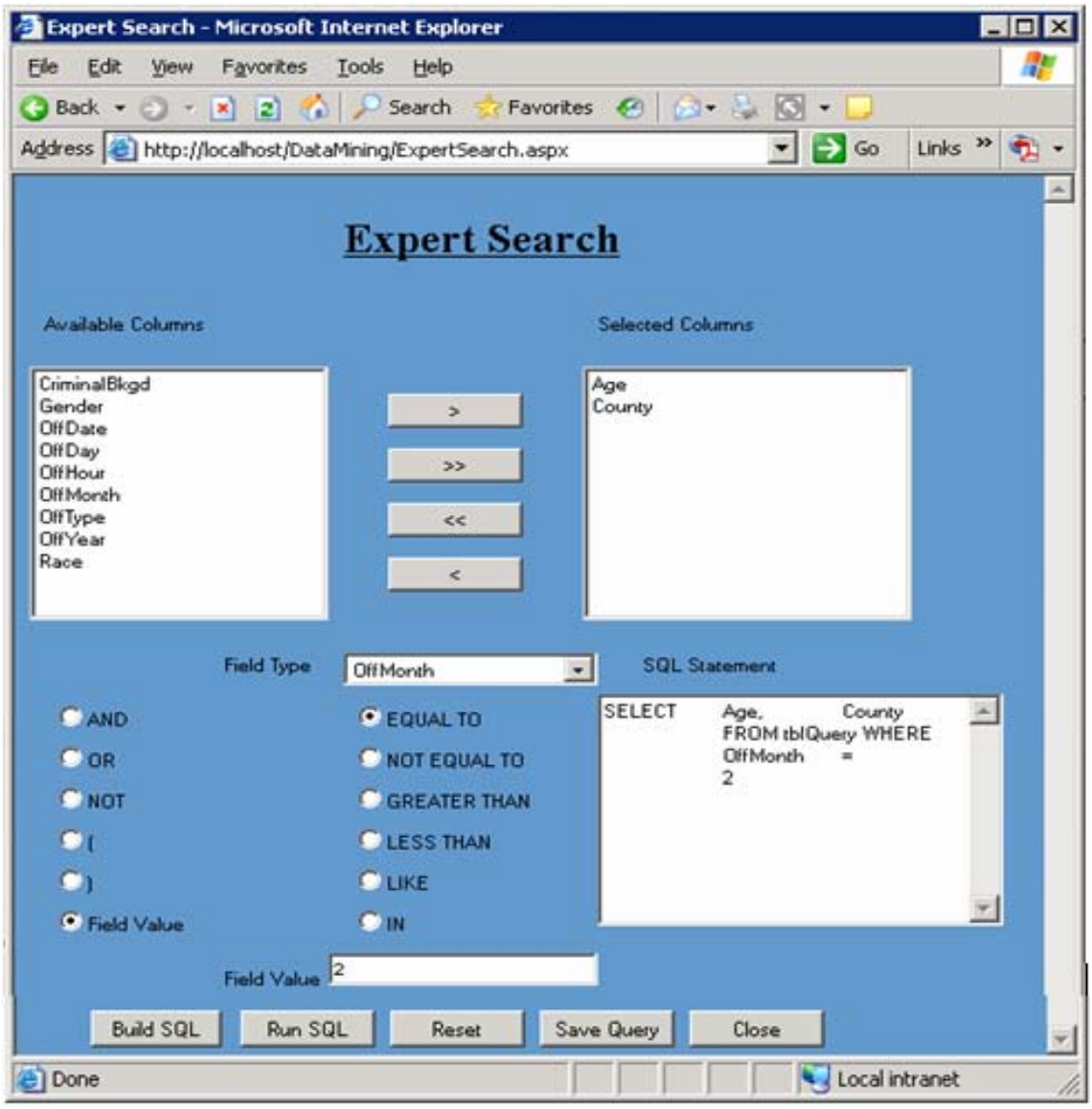

Figure 4.2: Expert Search

The query was built based on the choice made by the user with the “Build SQL” button and later run once the user clicks on “Run SQL” button. The outcome of the query put forth by the user, with only the selected fields, is show in Figure 4.3. Expert Search requires significant knowledge of SQL on the part of the user.

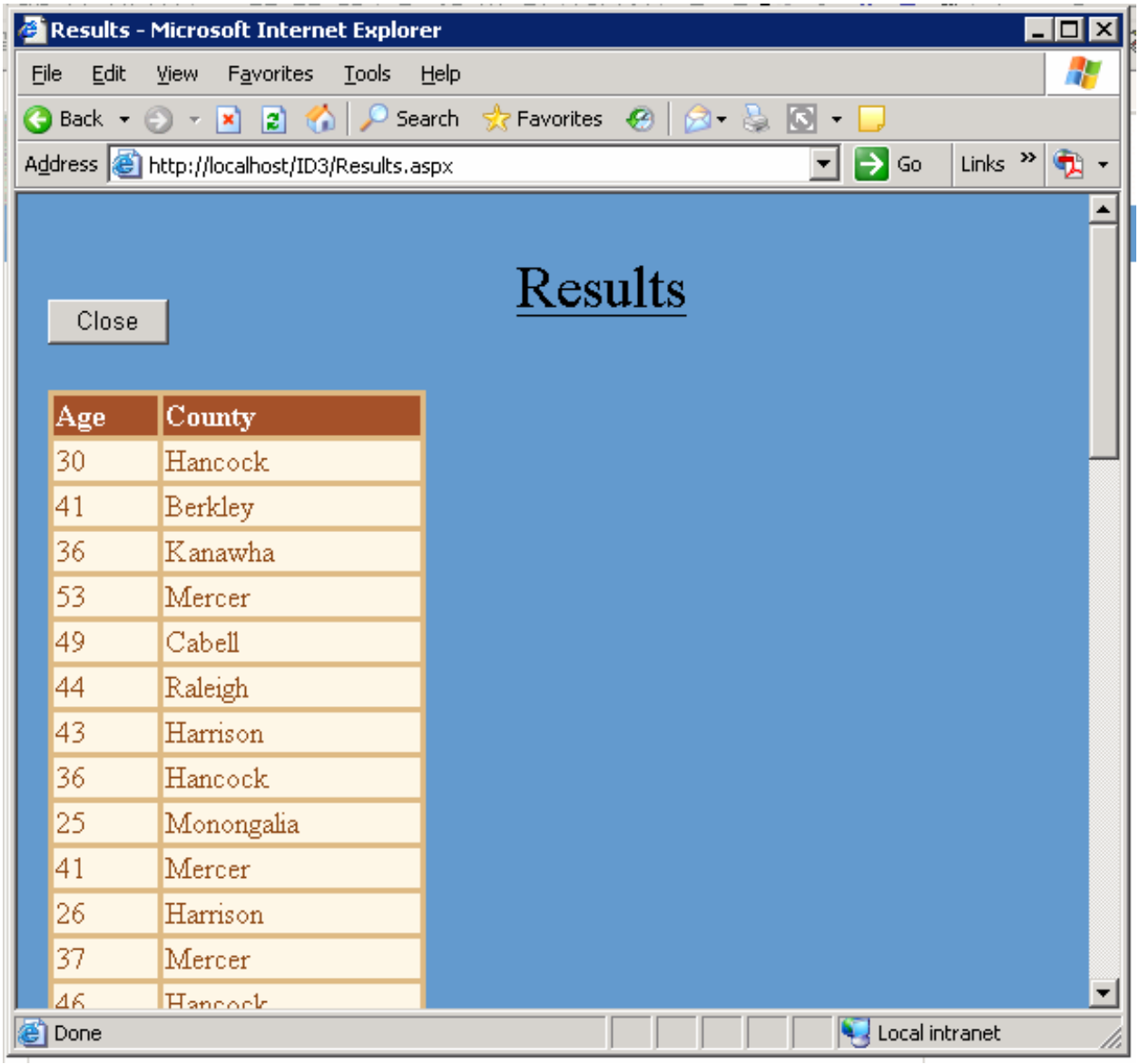

Figure 4.3: Result page of "Expert Search"

Still, few issues pertaining to querying the database remain unresolved. For nominal values, the user has $(2^{n-1} - 1)$ choices, where 'n' are the number of fields. That is, for 11 fields, the user would have 1023 choices to select from, without considering the "drop down" selections within each field or the conditions for each field. It is therefore overwhelming for the user without some initial leads. Moreover, the outcome was essentially a static picture and does not present the boundary conditions and / or over all picture to the user.

The above two issues are inherent to the searching method and there was no better method of querying the database. So, one has to look for other means of extracting information / knowledge from the data.

4.5 Data Mining: Algorithm Selection

The next level for information extraction is, moving from querying the data to mining it. One might recollect that, quite a few algorithms could be applied to mine the Query Table; as a matter of fact, most of the algorithms can mine the data. Thus, one must select appropriate algorithm, from the pool that was described earlier. However, the selection will depend on the type of data, the final objective and most importantly the kind of application of data, planned by the analyst.

From the Query Table (Section 3.7) it is known that the data are collection of nominal fields with no dependent variables or predefined groups. This rules out the Statistical and Neural Network based algorithms as they primarily work on training dataset with some sort of "dependent-independent" relationship to build the model. The Distance Based algorithms (Clustering) were difficult to apply and would further require significant changes to the existing data table to be applied effectively. This renders them practically infeasible.

Association Rules and Decision Rules / Trees show promise for our Data Mining application. These algorithms address the constraints put forth by the Query Table. Moreover, the outcome is directly seen by the end user and would not need further processing or interpretation.

Association Rules

The Association algorithm consists of three main functions, 1) Search Engine, 2) Item Set Generator, and 3) Pruning Algorithm. The Item Set Generator generates a combination of items from the given set. This is then used with help of Search Engine to search for similar patterns through the database. The combinations that have frequency below the desired level are pruned by the Pruning Algorithm. Moreover, the Item Set Generator and Pruning Algorithm work together such that all future instances of the combinations already pruned are automatically pruned. Thus, the solution contains a set of association rules derived from the dataset. One serious drawback of the Association technique is that they work only with nominal data, though this drawback can be overcome by using coding techniques.

For example, shown in Table 4.1 is a Level -1 set of items that are available in a store. These five items are being considered to find an association relation, if any, between them.

Table 4.1: Level-1 item set

Beer	Bread	Jelly	Milk	PeanutButter

The data available for the analysis are a transactional dataset as observed in Figure 4.2, where t1 to t5 and so on are the transactions. For each transaction there is set of items purchased from the store.

Table 4.2: Transaction Data

Transaction	**Item**
t1	Bread, Jelly, PeanutButter
t2	Bread, PeanutButter
t3	Bread, Milk, PeanutButter
t4	Beer, Bread
t5	Beer, Milk
:	:

In essence, the Association algorithm creates various combinations of the item set and runs a frequency check through the entire database. One has to specify a lower bound for the frequency below which the items sets would be pruned. The lower bound is

usually is terms of Minimum Support and Minimum Confidence. Based on this different combinations of items are pruned. Thus, one is left with the high frequency item combinations.

The Table 4.3 shows different combinations of the item or item sets that are being studied as well as the frequency of each combination. As explained one can select high frequency item combinations that will form the basis for Association Rules.

Table 4.3: Item set with frequencies

Set	**Frequency**	**Set**	**Frequency**
Beer	40	Beer, Bread, Jelly	0
Bread	80	Beer, Bread, Milk	0
Jelly	20	Beer, Bread, PeanutButter	0
Milk	40	Beer, Jelly, Milk	0
PeanutButter	60	Beer, Jelly, PeanutButter	0
Beer, Bread	20	Beer, Milk, PeanutButter	0
Beer, Jelly	0	Bread, Jelly, Milk	0
Beer, Milk	20	Bread, Jelly, PeanutButter	20
Beer, PeanutButter	0	Bread, Milk, PeanutButter	20
Bread, Jelly	20	Jelly, Milk, PeanutButter	0
Bread, Milk	20	Beer, Bread, Jelly, Milk	0
Bread, PeanutButter	60	Beer, Bread, Jelly, PeanutButter	0
Jelly, Milk	0	Beer, Bread, Milk, PeanutButter	0
Jelly, PeanutButter	20	Beer, Jelly, Milk, PeanutButter	0
Milk, PeanutButter	20	Bread, Jelly, Milk, PeanutButter	0
		Beer, Bread, Jelly, Milk, PeanutButter	0

Small item sets can be managed by simple logic or code. But the ones with large item sets require use of advance algorithms such as Apriori, CDA etc. Just as we saw a Level-1 item set, Table 4.4 shows a Level-2 item set where one can have combinations based on both field and field value.

Table 4.4: Level-2 item set

Type	**American**	**Italian**	**Mexican**	**Beverages**
ForHere	Burger	Pasta	Burrito	Soda
ToGo	Sandwich	Pizza	Quesadilla	Shakes
	Fries			Coffee

This table represents an actual scenario where Data Mining is useful for finding association rules. A logic similar to this was developed that can find association between different items but in paired form. Moreover, the complexity of algorithm was reduced by selecting only the first three field values having high frequency of occurrence.

Decision Rules

Decision rules, also called as Decision Trees, help with decisions that enable classification of a new unclassified dataset in an easily interpretable way. They are very easy and efficient way of mining the databases and work irrespective of the database size. They used to suffer from missing values and continuous data; which are now pruned and imputed. Also, most algorithms ignore correlations and interactions in data [2]. The leaves represent classifications, while branches represent conjunction sets of attributes that lead to classifications.

Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either every attribute has already been included along this path through the tree, or all elements associated with this leaf node have their score function (entropy) as zero.

For example, Table 4.5 shows set of factors that determine the Credit Score of an individual.

Table 4.5: Factors for determining credit score – Stage I

Credit Score	Bank Balance	Age	Sex	Job Title	Race
Average	> $10,000	> 35	Female	Entrepreneur	Asian
Bad	$10,000 > & > $5,000	35 > & > 24	Male	Manager	Black
Good	$5,000 >	24 >		Worker	Caucasian

Now, the Decision Rules algorithm finds out a Field and Filed Value combinations that maximizes the score function, which in our case can be Entropy or Gain, and that is the point of first split. Let's say from the dataset of 1000 entries, attribute Job Title = "Manager" has maximum gain of 700 entries, so is the point of first split; this will

generate two branches from the root which is the dataset based on condition of split. The new set of factors is shown in Table 4.6.

Table 4.6: Factors for determining credit score – Stage II

Credit Score	Bank Balance	Age	Sex	Race
Average	> $10,000	> 35	Female	Asian
Bad	$10,000 > & > $5,000	35 > & > 24	Male	Black
Good	$5,000 >	24 >		Caucasian

Let's say the next two splits are for Age = "> 35" and "Bank Balance = "> $10,000", with gain of 400 and 250 entries respectively, and, for Sex and Race the gain in zero. Thus, a tree with following conditions is generated.

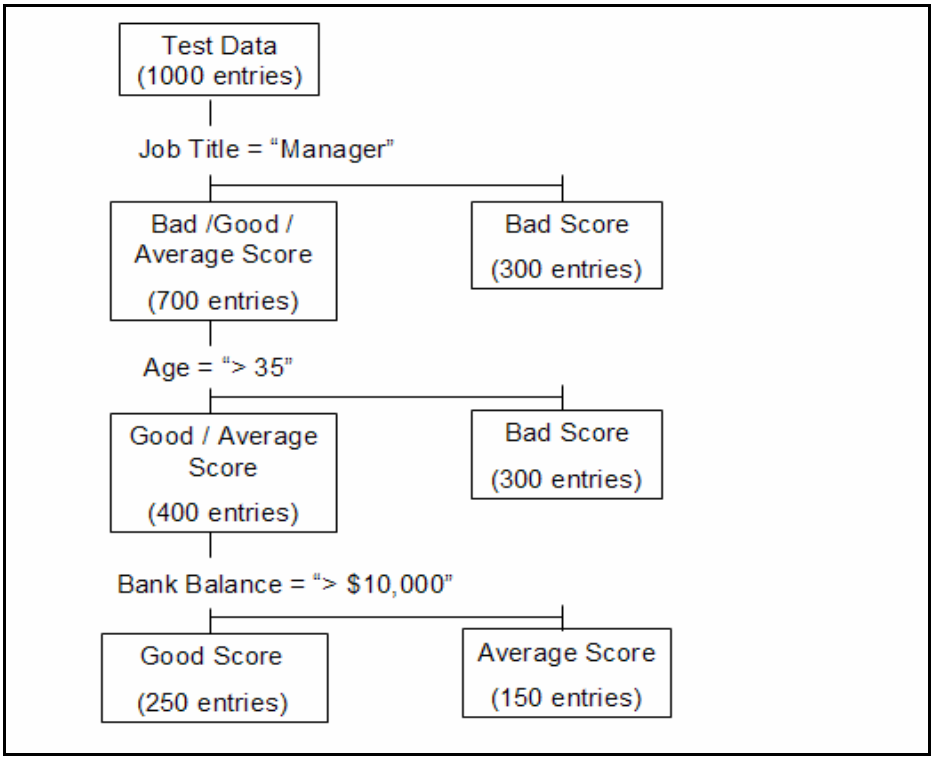

Figure 4.4: Good Credit Score Tree

This, tree is itself the conditions or decision rules for Good Credit Score. Decision rules similar to the example above were generated by implementing a parallel logic, the criteria for decisions being frequency of occurrence of a particular "Field-Value" combination.

It should be noted that one would use the algorithms in essence / logic along with DMQL concept (Section 2.7 – Page 30). All advanced or commercially available software use better algorithms and make use of Data Visualization to display the results.

4.6 Association Rules

In Association Rules, the user will have to select the fields from the database on which he wishes to mine the Query Table. For example, in Figure 4.5, the user selected "OffDay" (Offense Day) and "OffType" (Offense Type) to find associations. After pressing "Run" button, the generated output shows association between different days and offense along with frequency of occurrence.

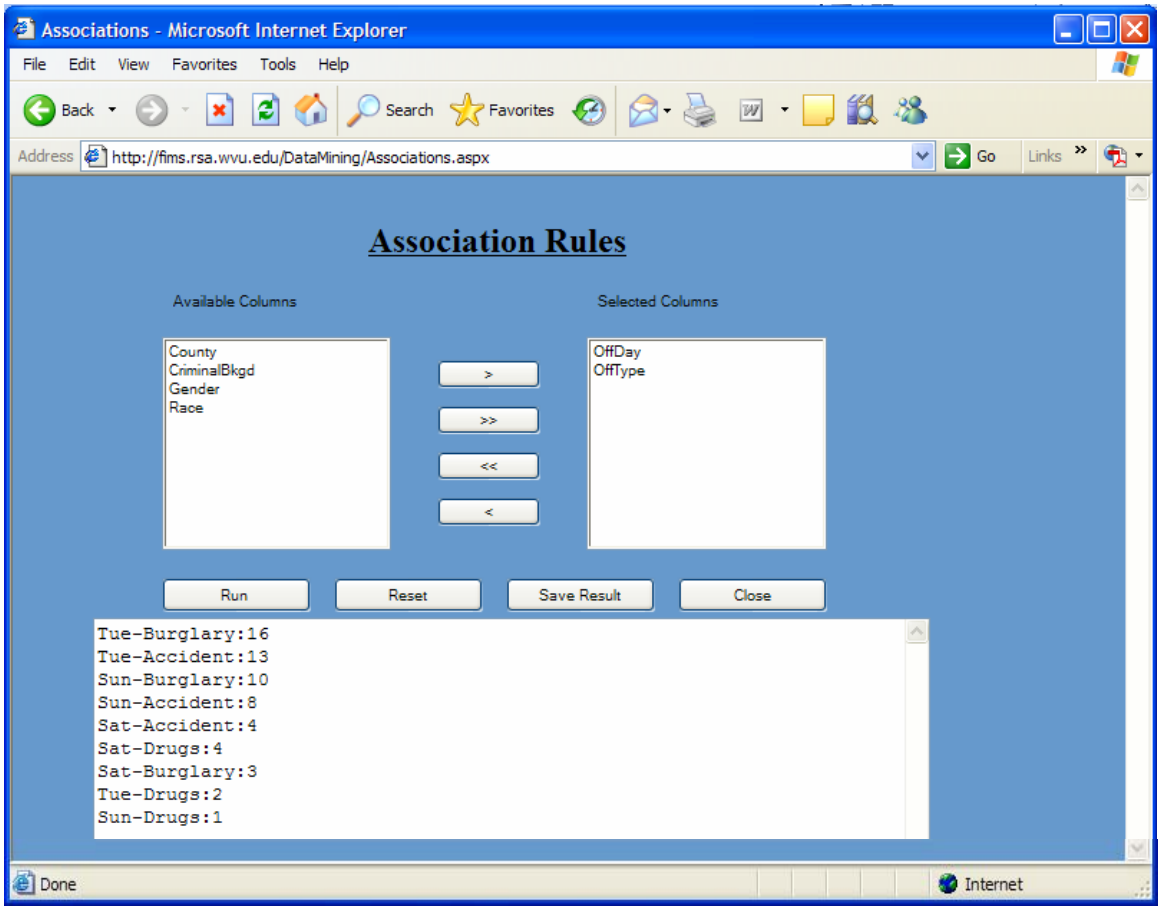

Figure 4.5: Association Rule

4.7 Decision Rules

The Decision Rules algorithm implementation is displayed in Figure 4.6, where, the user selected the Field and Field-value combination as “OffDay” and “Sat”. Also, he selected fields “OffType” and “Gender” to find out the Decision Rules. After pressing “Run” button, the generated output points out the frequency of occurrence particular offense and involvement of specific gender on Saturday. This kind of analysis can help in decision making in many real world scenario.

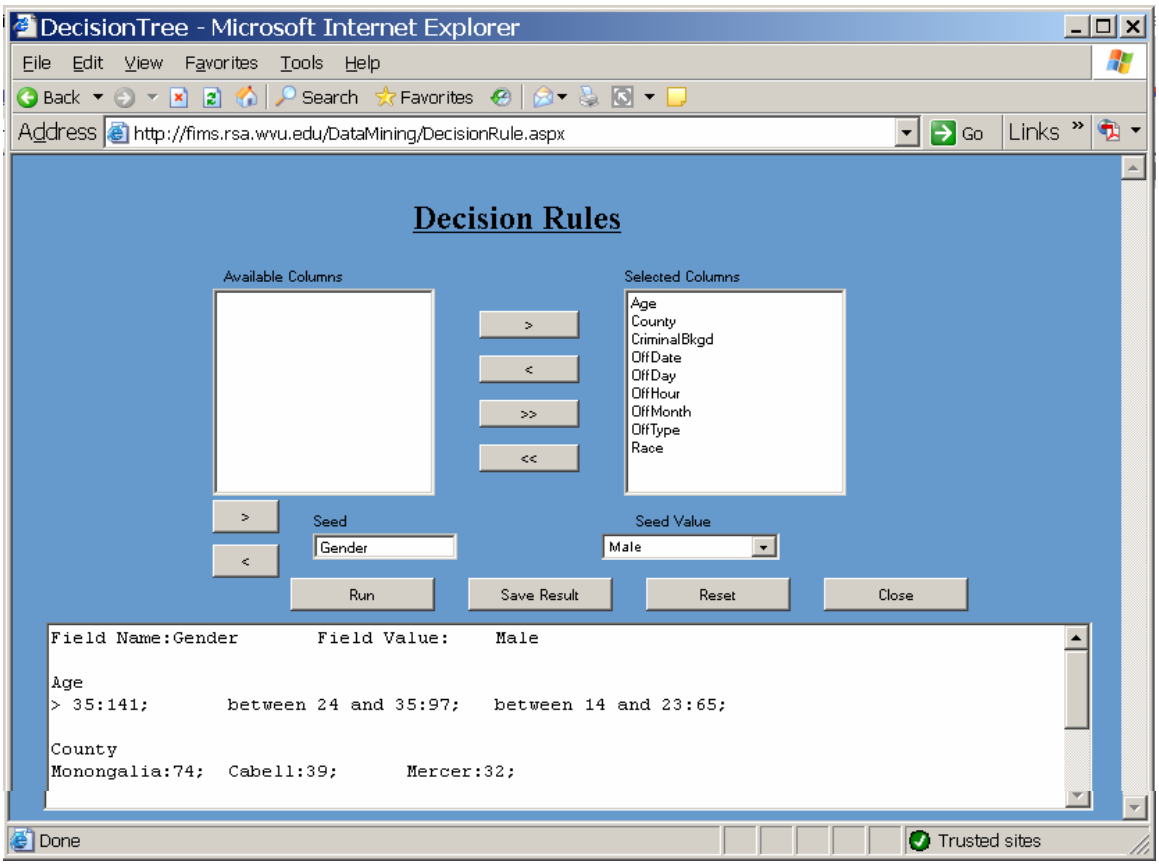

Figure 4.6: Decision Rule

4.8 Result Page

The results of all the techniques are summarized in the "Result Page", as displayed in Figure 4.7 below. This included storing the queries for the Searching procedures and the results for the Data Mining procedures respectively. This can be stored in the database by the user as per the requirement.

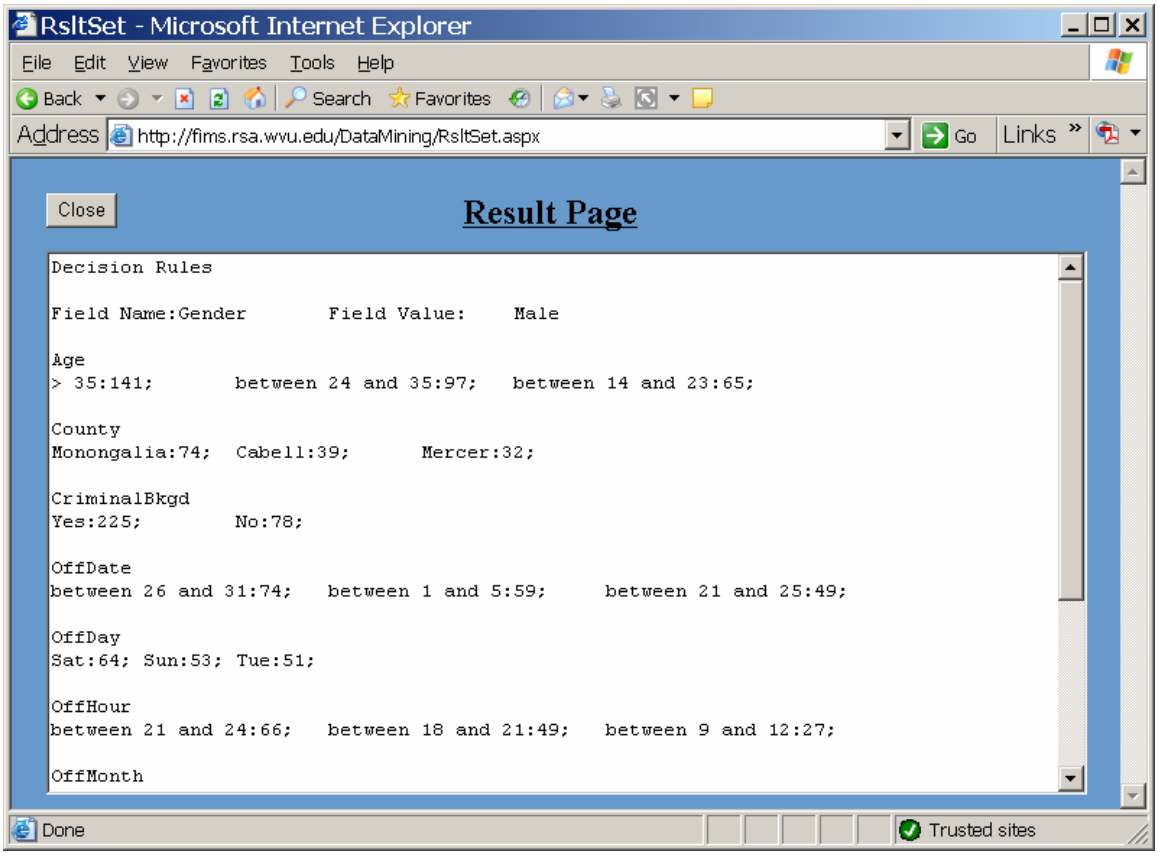

Figure 4.7: Result Page

4.9 Navigation Page

For the user to toggle between the various data search / mine operations navigation page, called “Data Mining Tools” was created. The user can click on a button and the user interface as discussed earlier is presented. Also, “Generate ARFF File” button is provided. This generates an ARFF file that will allow the user to use other Data Mining tools. The user interface for the same is shown in the Figure 4.8.

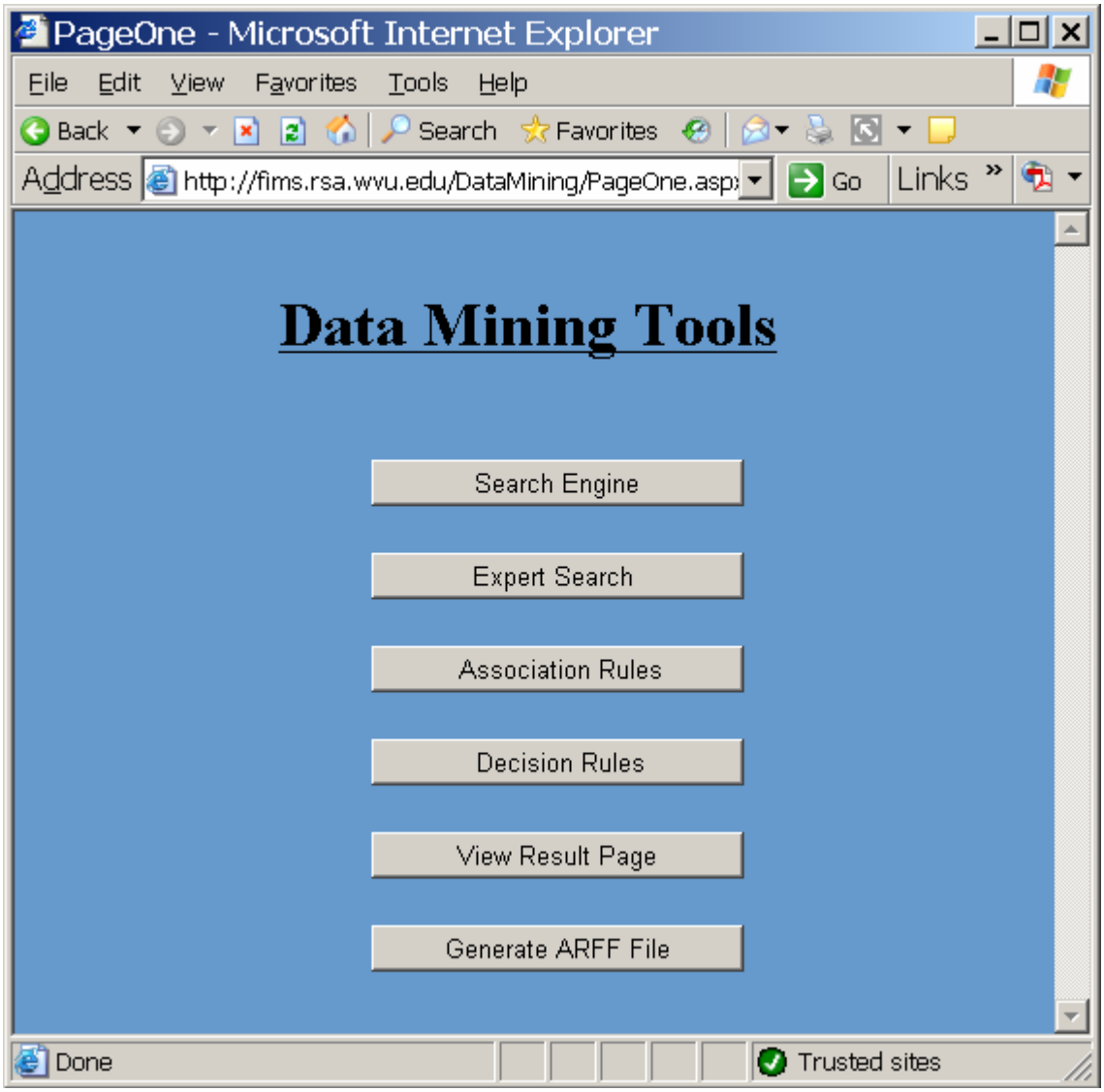

Figure 4.8: Navigation Page – “PageOne".

4.10 Other Data Mining Tools

WEKA is another prototype Data Mining tool available over the Internet. This is being developed by The University of Waikato, New Zealand [20]. Though it is implemented primarily in Java, recently many more computer languages have been added to it. WEKA is a shell command based program. Therefore it cannot be directly executed on the Web. The user has to create a file in Attribute Related File Format (ARFF) file as shown in Figure 4.9. The ARFF file can then be input to the WEKA program.

```
@relation TestData

@attribute ACN numeric
@attribute OffDate date "yyyy-mm-dd"
@attribute OffTime date "HH:mm:ss"
@attribute OffDay {Mon,Tue,Wed,Thru,Fri,Sat,Sun}
@attribute OffCounty {Berkley,Cabell,Hancock,Harrison,Kanawha,Marion,Mercer,Monongalia,Raleigh}
@attribute OffType {Accident,Arson,Assault,Burglary,Drugs,Murder,Rape,Sex_Assault,Weap_Threat}
@attribute Gender {M,F}
@attribute Race {White,Black,Hspnc,AorPI}
@attribute Age numeric
@attribute CriminalBkgd {Yes, No}

@data
1001,2004-1-2,12:28:00,Tue,Harrison,Assault,M,White,31,No
1002,2004-1-3,13:28:00,Wed,Cabell,Murder,M,Black,23,Yes
1003,2004-1-4,14:28:00,Thru,Monongalia,Drugs,F,AorPI,20,No
1004,2004-1-10,15:28:00,Sat,Marion,Burglary,M,White,24,Yes
1005,2004-1-10,16:28:00,Sun,Hancock,Murder,M,White,36,Yes
1006,2004-1-13,17:28:00,Tue,Berkley,Accident,F,AorPI,52,No
1007,2004-1-16,18:28:00,Fri,Monongalia,Sex_Assault,M ,White,40,Yes
```

Figure 4.9: Test Data in ARFF.

The above approach provides a generalized way to study the database using variety of algorithms.

The server based model has button “WEKA” that generates an ARFF file, called “TestData.arff”, dynamically from the query table “tblQuery”. This will also open the WEKA front page as in Figure 4.10.

Figure 4.10: WEKA – Front Page

The user has to click the button named “Explorer” and then open the file by just clicking the “Open File” button, as in Figure 4.11 and select the “TestData.arff” file.

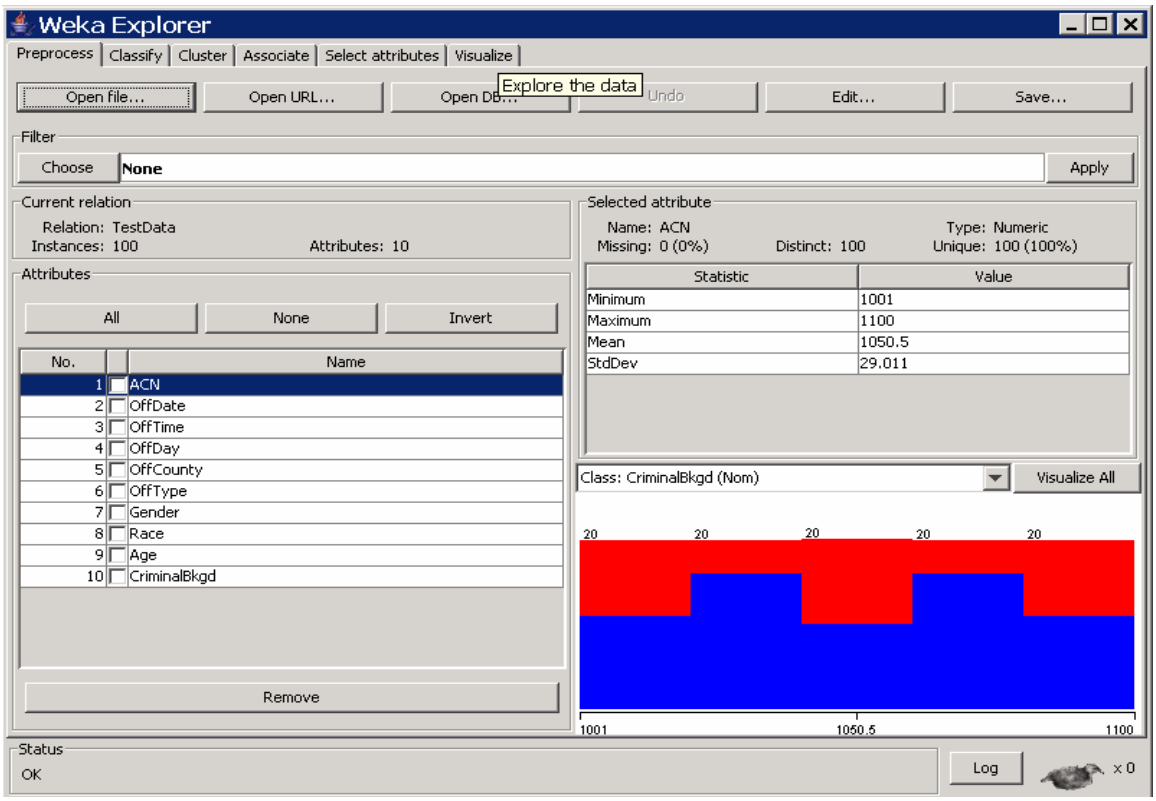

Figure 4.11: WEKA - Explorer Page

Further, the user can choose the required fields and click on “Visualize” tab at the top of the application. He can view the confusion matrix of the data, as shown in Figure 4.12.

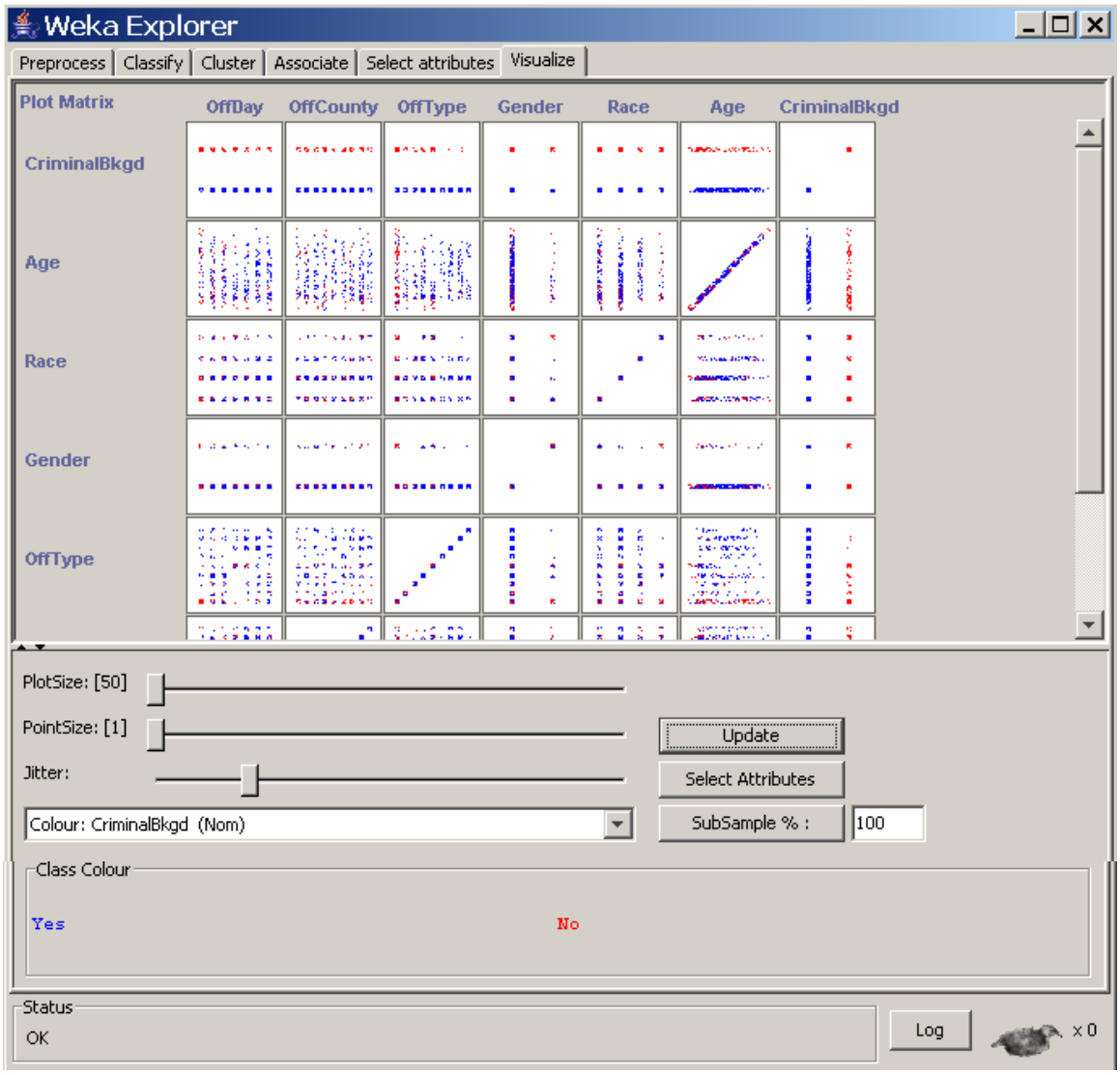

Figure 4.12: WEKA – Visualize (Confusion Matrix)

By clicking on one of the graphs and by changing the x-axis, y-axis and color options one can view different trends and patterns in the data as in Figure 4.13. The jitter scale on the plot should be at maximum for best visualization.

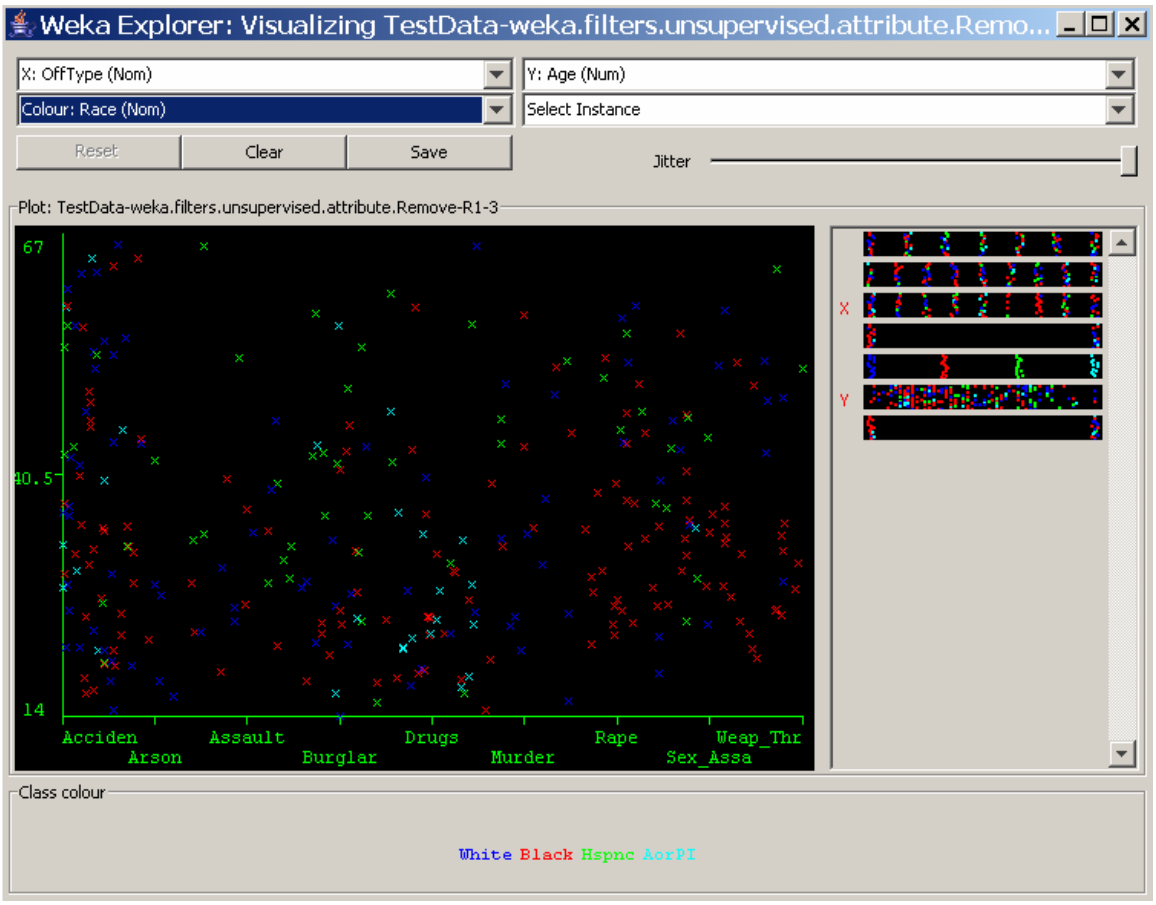

Figure 4.13: WEKA – Plot from confusion matrix

5. APPLICATION STUDY

Data Mining Methodology was briefly described in Section 2.2. This chapter describes how that methodology was carried out for an application study. Similar to Data Mining methodology, the application study was also done in a phased manner.

5.1 Business Understanding Phase

The primary objective is to give to the end user (e.g., a crime investigator) a simple tool that will help him or her utilize the crime data to identify criminals and crime patterns. Even with minimal Data Mining knowledge one should be able to use the tool and generate meaningful results. Also, the tool should provide Web functionality so that the user can access it from anywhere and at any time.

Thus, the user will be provided with Web-based basic and advanced search techniques to query for specific inputs. Additionally, he or she will be offered suitable Data Mining techniques to generate vital information (e.g., crime-related leads) from the data. To visualize the data, he or she will use the WEKA-Experimenter.

5.2 Data Understanding Phase

The data required for Data Mining were not available due to security reasons, so major data fields used in the crime database were determined from the Web and other sources. However, a need for data (query) table for Querying and Data Mining was identified.

The data consisted of nominal, non-ordinal and/or null values with varied data types such as numeric, datetime, and varchar. It did not have "Dependent – Inependent" relationship nor was it available in any particular sequence. Moreover, the data contained

13 fields with a maximum of 112 (crime types) possible values for a given field. Such data posed unique constraints concerning the selection of Data Mining techniques.

5.3 Data Preparation Phase

To develop the Data Mining Tool, the crime database was created, as described in Chapter 3, and populated. Initially, the "OffDate" and the "OffTime" fields of the tables were populated with the help of random numbers and the "OffDay" field was populated with the help of calendar. From the "Most Wanted" criminal data of West Virginia Police Dept. [21] and Bureau of Justice Statistics database [22], other fields such as "OffType," "Gender," "Race," and "CriminalBkgd" were populated. But these data was distributed in three tables "dtOffense," "dtSuspect_Physical," and "dtSuspect_Variable."

For ease of Data Mining application, the data was put into single data table called "tblQuery." This was done in two stages by selecting data from the database in middle tier and inserting the results back to the query table "tblQuery." The test data that were populated can be seen in Appendix A. Finally, the query table "tblQuery" was used for the Data Mining tool.

5.4 Modeling Phase

As described in Chapter 4, the searching algorithms as well as appropriate Data Mining techniques were implemented according to the established objective. This was done in ASP.NET to give Web-based functionality to the tool. But this tool lacked visual depiction of the results. To overcome this issue, the user was given an option to use the data with WEKA software at server end. Another aspect concerning the tool was that the Data Mining techniques used hard coded field descriptions to work with data, as it was necessary to reduce the complexity the algorithm.

5.5 Evaluation Phase

The model was evaluated in two stages. At first, the model was run directly on the given data. The outcomes of all the techniques were directly validated by MS Excel and SQL Query Analyzer by running specific queries. The validation results were judged by actual numeric value of the output of the techniques. This ensured that the values generated were correct and the relation that is expressed is true. The second stage was to test the model against an external source. It was decided to test results of the model against WEKA and then compare the output. This testing was carried out, but the actual numeric values could not be confirmed, as WEKA does not report an outcome in a form similar to ones tool. The relationships represented by WEKA and the tool were found to be same during testing phase. Moreover, the code for dynamic conversion of test data into ARFF was checked by comparing file created by code with the actual physically created file.

5.6 Deployment Phase

This is the last stage of the Data Mining Process. Once the database and software was implemented and tested, the tool was put through a pilot run and further deployment. Often, steps (2) to (5) of the methodology are re-run for further refining the Data Mining Tool. For an application study, some patterns were embedded into the test data. In this study a pattern refers to relationship between data fields that have some relevance to crime. These patterns resemble actual patterns in a Data Mining scenario. The patterns that were embedded are:

a. Harrison County had a high rate of accidents.
b. Crime against women occurred mostly on weekends and was committed by older age group.
c. Drugs were consumed more during weekends by young men.

d. Monongalia County had significant number of burglary and drug cases.

These patterns were placed in a very non-obvious way. A very low frequency of occurrence was maintained, such that the inherent noise in the data may overcome the intended patterns and one could get an altogether different and non-obvious pattern similar to real life scenario. An attempt was made to identify these patterns with the Data Mining Tool.

Figure 5.1 shows the implementation of the Deployment Phase. The Validation of the tool is only done initially during the pilot of the tool. Once implemented the user will select the tools based on the task. For Searching and Querying operation he will use the "Search Engine" and "Expert Search"; while to mine these data for non-obvious crime patterns the "Association Rules" and "Decision Rules" will be used. User may use searching tools to further query these data based on the mining results.

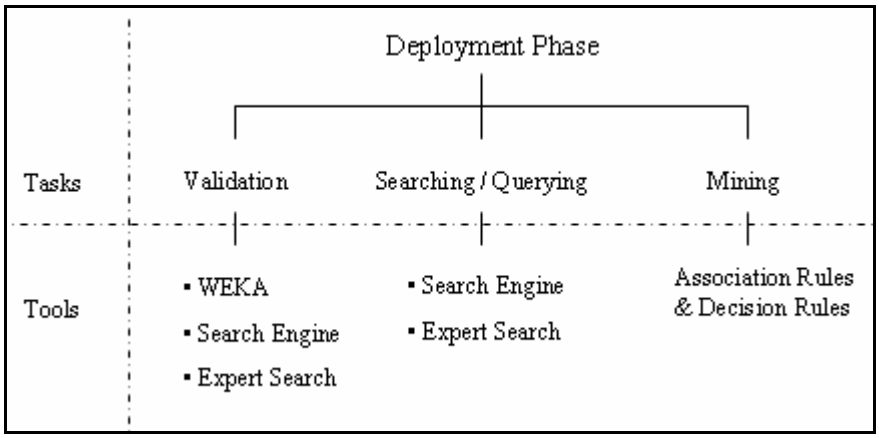

Figure 5.1: Tree for Deployment Phase

Data Mining with Searching Tools

The outcomes of the Search Engine were unable to identify any patterns in the data. The Expert Search also worked on similar lines, as initially there were no leads available. One faced an overwhelming number of combinations while using the two searching tools.

When the patterns were specifically queried, even then the Search Engine was unable to clearly determine the conditions that govern them. While the Expert Search with a better interface could find the results, they were isolated or exact outcomes and did not clearly determine if they represent any patterns.

Data Mining with Association Rules

The Association Rules was used with selection of various combinations of the four fields – "OffDay," "OffType," "OffCounty," and "Race." They were:

a. OffDay – OffType – OffCounty – Race
b. OffDay – OffType – OffCounty
c. OffDay – OffType
d. OffDay – OffCounty

These combinations were randomly chosen. The other two field types, "CrimeBkgd" and "Gender," of the query table "tblQuery" were left out of the study as they had only two Field Values. By selecting them, the outcome would be biased, as these field values will split the data into two and always end up with higher frequency. Also, the numeric fields, such as "Age," "Date," and "Time" of Crime are excluded as they cannot be used for the Association Rules technique in their present form and coding them requires significant knowledge and input that was beyond the scope of the thesis.

The Association Rules technique selected the first three Field Values from each of the above Fields, with a high frequency of occurrence, and checked their combinations to find out associations in the data. The four fields "OffDay," "OffType," "County," and "Race" are selected by choosing each and clicking the right arrow key. When the "run" button is pressed, the outcome of association for the selected fields is displayed in the text box, as show in Figure 5.2. The complete outcome of the above textbox is displayed in Figure 5.3, while the outcomes for different field combinations are shown in Appendix C.

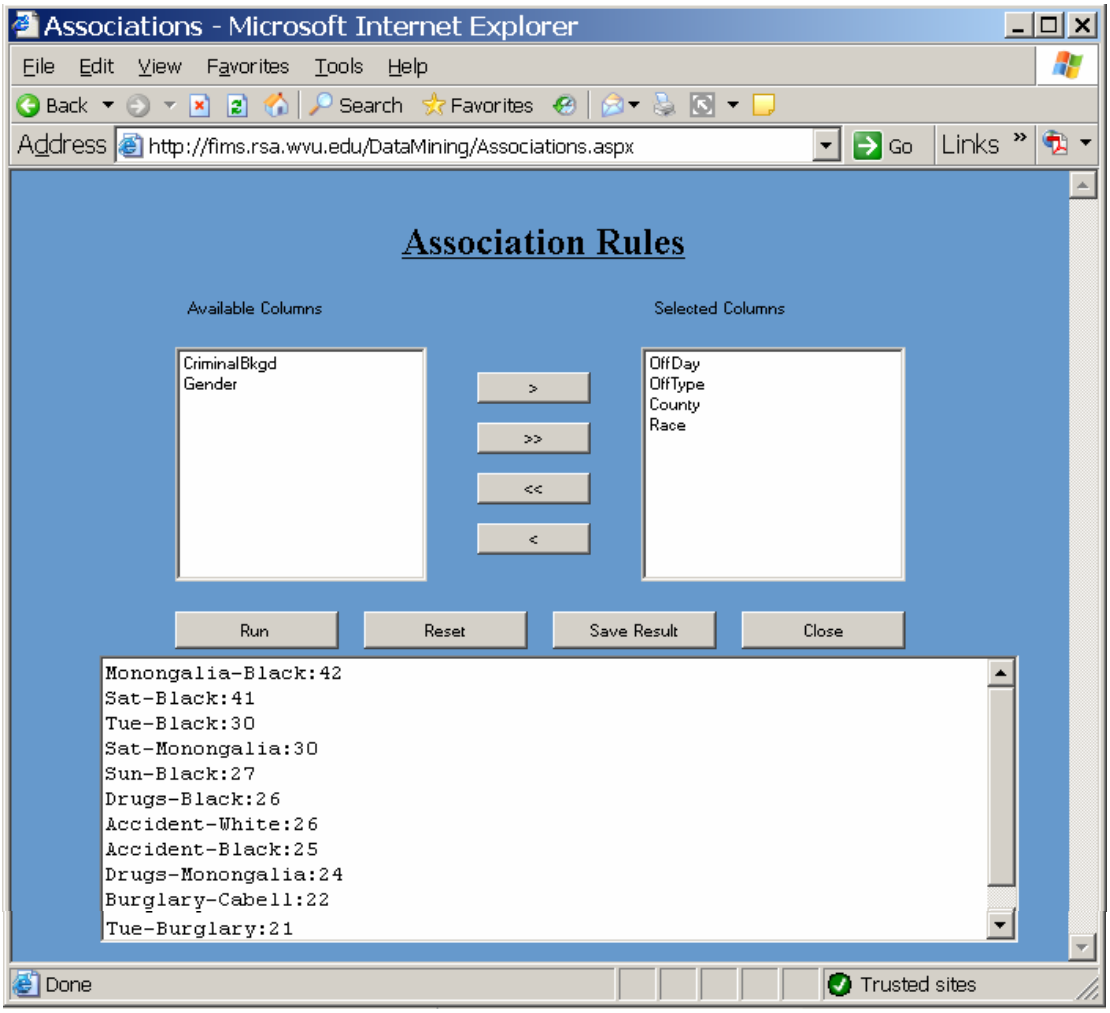

Figure 5.2: Association Rules with selected fields

Monongalia – Black	: 42
Sat – Black	: 41
Tue – Black	: 30
Sat – Monongalia	: 30
Sun – Black	: 27
Accident – White	: 26
Drugs – Black	: 26
Accident – Black	: 25
Monongalia – Drugs	: 24
Cabell – Burglary	: 22
Tue – Burglary	: 21

Figure 5.3: Association Rules outcome

The selective (top four associations) view of the Association Rules for various combinations is shown in Table 5.1.

Table 5.1: Tabulated Outcome of Association Rule

Combinations	Selective Outcome
"OffDay," "OffType," "County," and "Race"	Monongalia – Black
	Sat – Black
	Tue – Black
	Sat – Monongalia
"OffDay," "OffType," and "County"	Sat – Monongalia
	Drugs – Monongalia
	Tue – Burglary
	Burglary – Cabell
"OffDay" and "OffType"	Tue – Burglary
	Tue – Accident
	Sat –Drugs
	Sun – Burglary
"OffDay" and "County"	Sat – Monongalia
	Tue – Cabell
	Tue – Monongalia
	Sun – Cabell

From the table, it was observed that most of the incidents in the data that have high association occur on Tuesdays and/or Weekends. For the first combination, "OffDay – OffType – County - Race," it was seen that the associations mainly included just the Black race. The combination of "OffDay – OffType – County" showed association between Tuesday and Burglary, Saturday and Monongalia, Drugs and Monongalia, and Burglary and Cabell. The combination of "OffDay – OffType" showed associations on similar lines involving Tuesday with Burglary and Accident, Saturday and Drugs and Sunday and Burglary. While in the last combination, "OffDay – County," it was seen that Monongalia County was associated with Tuesday and Saturday, while Cabell was associated with Tuesday and Sunday.

Data Mining with Decision Rules

The Decision Rules followed the Association as the Data Mining Tool. The field "OffType" was randomly selected and all the offenses or field values were selected one at a time. This "Field-Field Value" combination was then worked against other fields that were selected for generating the Decision Rules. The Figure 5.4 shows the outcome of the Decision Rules for Field Name = "OffType" and Field Value = "Accident."

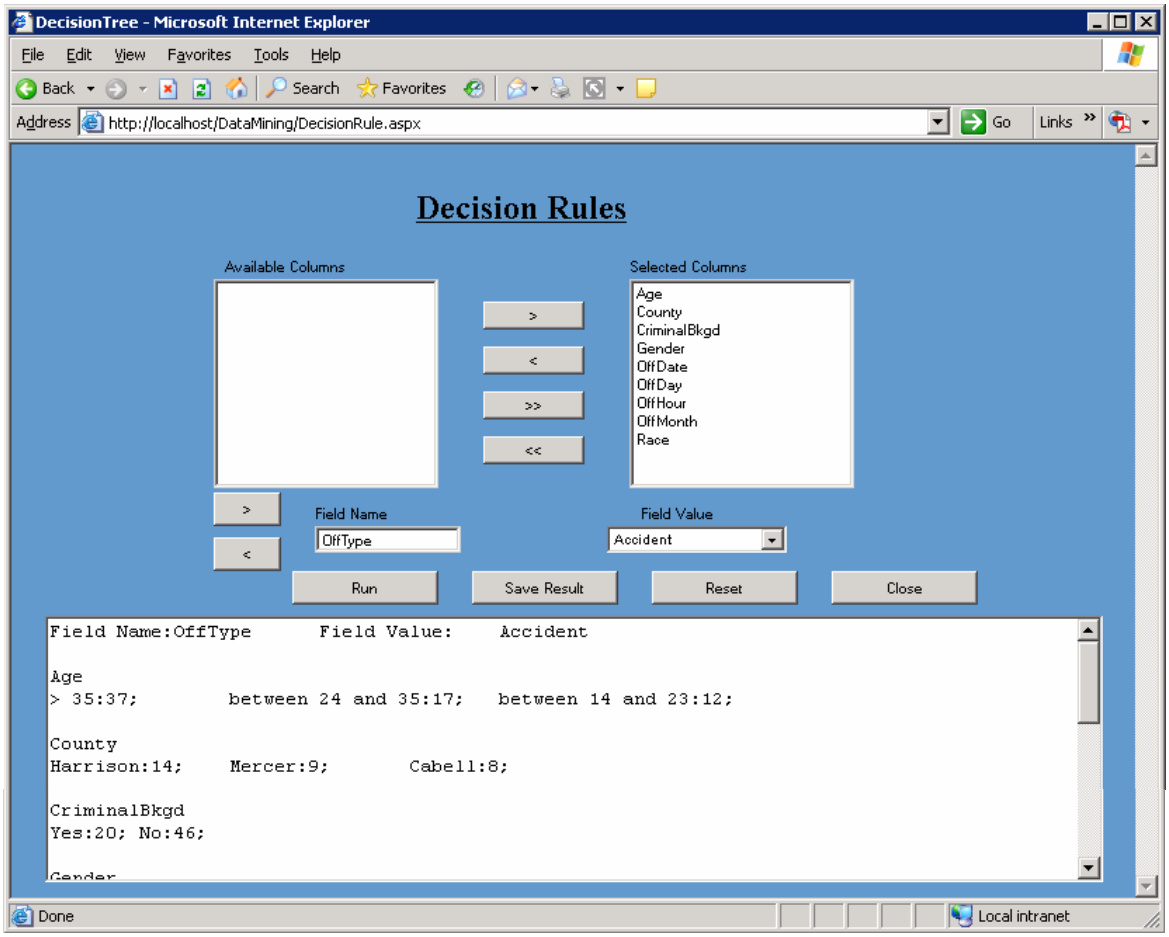

Figure 5.4: Decision Rules for OffType = "Accident"

The complete outcome of the above textbox is displayed in Figure 5.5, while the outcomes for different field combinations are shown in Appendix D.

Field Name: OffType Field Value: Accident

Age
> 35: 37; between 24 and 35: 17; between 14 and 23: 12;

County
Harrison: 14; Mercer: 9; Cabell: 8;

CriminalBkgd
Yes: 20; No: 46;

Gender
Male: 56; Female: 10;

OffDate
between 6 and 10: 18; between 26 and 31: 16; between 1 and 5: 11;

OffDay
Mon: 25; Wed: 14; Tue: 13;

OffHour
between 6 and 9: 18; between 15 and 18: 9; between 9 and 12: 8;

OffMonth
Jun: 10; Jul: 9; Sep: 6;

Race
White: 26; Black: 25; AorPI: 8;

Figure 5.5: Decision Rules outcome

The technique finds the maximum frequency of occurrence of all field values of the selected fields with the Field-Field Value combination from the data table "tblQuery." Then, the top three occurrences for each selected field were listed as outcome of the Decision Rule. The outcomes for each Field Value or Offense Type were shown in Appendix D. The selective view of the outcomes, with significantly high occurrences from the group, is shown in Table 5.2.

Table 5.2: Tabulated Outcome of Decision Rule	
Offense Types	**Selective Outcome**
Accident	Age > 35, County - Harrison, Day – Monday, Race - White & Black
Arson	Age between 24 & 35, County - Harrison, Day – Sunday, Race - Black
Assault	Age between 24 & 35, Day – Sunday, Race - Black
Burglary	Age >24, County – Cabell, Day - Tuesday
Drugs	Age between 14 & 23, County - Monongalia, Day – Friday & Saturday, Race - Black
Murder	Age > 35, County - Marion, Day – Sunday, Race - Black
Rape	Age > 35, County - Monongalia, Day – Saturday, Race, Black
Sex_Assault	Age > 35, County - Monongalia, Day – Saturday, Race - Black
Weap_Threat	Age > 35 and between 24 & 35, County - Hancock, Day – Tuesday & Sunday, Race - Black

From the Table 5.2, the Decision Rules for each of the Offense types were observed. Most Crimes against women occurred on weekends and the offenders belonged to higher age group (35 to 45). Based on Age groups, namely between 14 & 23, between 24 & 35 and greater than 35, the offense type changed from Drugs, Arson and Assault to more serious ones such as Murder and Weapons Threat. Based on Race for most of the crime incidents, the Black population had maximum involvement while there are Asian or Pacific Islanders who were involved in Drugs-related offenses. While considering Day of offense, it was seen that Drugs offense and Crime against Women, were reported more, in the later half of the week, on Friday and Saturday respectively. Also, it was seen that on Monday, a greater number of Accidents occurred, while more Burglary incidents were reported on Tuesday. With County of Offense, it was observed that Monongalia County reported a greater number of Drug and Crime against Women incidents. Harrison County lead Accident- and Arson-related cases and finally Burglary, Murder and Weapon Threat were reported in Cabell, Marion, and Hancock counties respectively.

Interpretation of Data Mining Tools

In the above sections, the outcomes of the two Data Mining Techniques were gathered in a selective manner. Those outcomes are interpreted in this section. The approach is to first compare the results of both the techniques and then figure out whether the selected Field Values indeed turn out to be a pattern. Moreover, the Decision Rules give us Age related outcomes, so those would not be compared but directly used for interpretation.

The Association Rules, from Table 5.2, showed only the Black and White race for crime. This can be attributed to the Test Data where 77.67% of the population was represented by these two races. From other combinations the ones with common Field Values were paired. For this the Field with more occurrences was selected as common field, which as per the Association outcome was Day of Offense. These pairs (links) are show below:

a. Tuesday – Burglary and Accident
   – Cabell and Monongalia
   – Black
b. Saturday – Monongalia
   – Black
c. Sunday – Burglary and Accident
   – Monongalia

There was also another single pair:

d. Drugs – Monongalia

The pairs identified above were then compared with the outcomes of Table 5.2. The commonalities found by comparison were:

A. Drugs – Monongalia – Saturday – Black
B. Burglary – Cabell – Tuesday

Moreover, from the Table 5.2, by pairing based on the outcomes, another single pair was found:

C. Age Between 24 & 35 – Sunday – Black – Assault
– Weap_Threat

D. Age > 35 – Monongalia – Saturday – Black – Sex_Assault
– Rape

The last four selected cases (from A to D) were then run specifically through the searching tools and their number of occurrence was reported in the Table 5.3.

Table 5.3: Number of Occurrences of Combinations

S. No.	Combinations	No of Occurrence
1	Drugs – Monongalia – Saturday – Black	8
2	Burglary – Cabell – Tuesday	5
3	Age Between 24 & 35 – Sunday – Black – Assault	5
4	Age Between 24 & 35 – Sunday – Black – Weap_Threat	2
5	Age > 35 – Monongalia – Saturday – Black – Sex_Assault	4
6	Age > 35 – Monongalia – Saturday – Black – Rape	6

It was observed that for the Drugs offense had the highest number of occurrences followed by Rape, Burglary, and Assault. The Weapons Threat reported just two incidents, so they would be pruned.

The outcome was then compared with the pattern (Page 67) that was planted into the test data of the query table "tblQuery." The Data Mining tool was not able to find the sub-pattern (a). This was selected in Table 5.2 but was not selected for further study as it neither match the relations in Table 5.1 nor formed pair in the same table. Also, the sub-patterns (b) and (c) were identified by the Data Mining tool. For pattern (d), the tool could find the relation for the Drugs offense, but for the Burglary offense no relation was identified. Instead, the tool found a new relation shown in Table 5.3 – S. No. 2. These results were validated with the WEKA software.

Data Mining with WEKA

The test data from table "tblQuery" were converted into ARFF format with a code specifically developed to dynamically access data from the database and convert them into an ARFF file. Thus, the generated file was input with WEKA software. The data were preprocessed with an unsupervised filter "Discretize –B 10 –M -1.0 –Rfirst –last" and then "Apriori" and "Tertius" classifiers (Associators) were used. Figure 5.6 shows the outcome for WEKA with Discretize filter and Tertius classifies with 6 attributes.

```
Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-weka.filters
Instances:    330
Attributes:   6
              OffDay
              OffCounty
              OffType
              Gender
              Race
              Age
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.347445 0.042424 */ OffType = Drugs ==> OffDay = Thru or Race = AorPI or Age = '(-inf-19.3]'
 2. /* 0.326950 0.030303 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Race = AorPI
 3. /* 0.320195 0.033333 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Age = '(-inf-19.3]'
 4. /* 0.319129 0.057576 */ OffType = Drugs ==> OffDay = Thru or Age = '(-inf-19.3]'
 5. /* 0.317121 0.048485 */ OffType = Drugs ==> OffDay = Thru or Gender = F or Age = '(-inf-19.3]'
 6. /* 0.317032 0.057576 */ OffCounty = Monongalia ==> OffDay = Sat or OffType = Sex_Assault or Age = '(19.3-24.6]'
 7. /* 0.312442 0.039394 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia
 8. /* 0.310521 0.027273 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Age = '(24.6-29.9]'
 9. /* 0.309606 0.033333 */ OffType = Burglary ==> OffDay = Wed or OffCounty = Cabell or Race = Hspnc
10. /* 0.303918 0.033333 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Gender = F

Number of hypotheses considered: 175531
Number of hypotheses explored: 87890
Time: 01 min 01 s 515 ms
```

Figure 5.6: Tertius classifier with 6 attributes using Discretize filter

The outcomes for all other combinations of these classifiers and number of attributes are shown in the Appendix E. The selective outcomes are listed by pairing them, as done in prior sections, in the Table 5.4.

Table 5.4: Tabulated outcome of WEKA				
S. No	**Apriori**			
1	Saturday	Black	Male	
2	Monongalia	Black	Male	
	Tertius			
3	Accident	Harrison	Monday	
4	Accident	Harrison	Female	
5	AorPI	Accident	Harrison	
6	Drugs	Thru	Monongalia	Age < 19
7	Monongalia	Saturday	Sex_Assault	Age between 24 & 29
8	Burglary	Tuesday	Cabell	Hspnc
9	Burglary	Cabell	Hspnc	
10	Drugs	Thursday	AorPI	

The WEKA software identified all the patterns that were planted into the database. Moreover, the newly discovered relation that was identified by the Data Mining tool was also identified by the WEKA software in Table 5.4 – S.No. 8 and 9. From relation (d) of the planted pattern (Page 67), even WEKA did not find any relation between Burglary and Monongalia County. The WEKA outcomes had many more paired relations that included the "CriminalBkgd" and "Gender" fields. These relations were ignored as explained in previous section. In cases of the Drug offense, the Data Mining model identified the Day of offense in the link as "Saturday," while WEKA reported it to be "Thursday." The database was specifically queried for this and it was found that number of occurrence of Drug offense on "Thursday" and "Saturday" was 10 and 8 respectively. This was due to the additional constraint, Race = Black, that was included while using Data Mining tool. The other small variations in the outcome of the Data Mining and WEKA software can be justified as both have different approaches for finding the patterns. The WEKA also found one more relation in the database that is listed in Table 5.4 – S.No.10. When specifically queried it was found that Asian or Pacific Islanders were involved in Drugs offense significantly.

The outcome of WEKA validates the Data Mining tool (stage II) as discussed in the Evaluation Phase (Page 67). Further, one can make use of WEKA to "Visualize" the patterns and other interesting combinations as described in Section 4.10.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

The objective of this study was to develop a Data Mining tool for the crime investigator. The tool should be able to identify obvious and non-obvious crime patterns in the database that would be helpful in investigations. The tool needed development of a crime database as well as the software to work on the data. Moreover, the software included development and implementation of data searching and mining tools. The tool was developed and implemented using the MS SQL Server and ASP.NET. The basic and advanced search techniques as well as the Association and Decision Rules techniques were implemented successfully. A suitable user interface was created to access them and a navigation page was provided to use different tools preferentially.

For validation, the tool was first checked by actual querying the database with SQL Query Analyzer. The test data were created with a pattern planted deliberately in it. The outcome of the tool for the test data was then compared with the outcome of WEKA software. It was observed that the tool was able to find the non-obvious patterns in the test data. Some other features of the tools are:

a. The Data Mining tool is customized for the application as per the objective; so unlike WEKA, which is a generalized application, it requires no preprocessing of the data.

b. The outcome of the tool showed direct relations in the field values and was easily interpreted even with limited knowledge of the Data Mining tools.

c. Further, the outcome can be immediately analyzed by using specific queries with the searching techniques provided in the tool. This required basic SQL querying knowledge on part of user.

d. The tool can be used over the Web and, at the same time, ensure the privacy and security of the data.

e. An ARFF file can be generated, which acts as a cross platform, thus enabling other Data Mining software to use the data for analysis.

6.2 Future Work

The tool provided satisfactory results and was able to determine crime patterns from the data for further investigation. This satisfies the foremost objective of the tool. However, there are several improvements that can be made. Some of these are listed below:

a. Currently, the tool does not have an ability to visualize the data. This makes it dependent on other applications such as WEKA for the same. Data Visualization could be incorporated in the tool.

b. The Data Mining algorithms were specific to the application. These algorithms can be further generalized so that they can be used in other cases and applications.

c. The query table "tblQuery" was essentially static in terms of fields and field types. An application can be built to achieve dynamic generation of the table based on user choice to make the tool more robust.

d. Further, an option to generate an XML file from the SQL data, similar to an ARFF file could be added. This would provide a general purpose platform for use of other Data Mining techniques.

REFERENCES

[1] Daniel T. Larose, “Discovering Knowledge in Data”; Wiley-Interscience, ISBN: 0-471-66657-2.

[2] Margaret H. Dunham, “Data Mining: Introductory and Advanced Topics”; Prentice Hall, ISBN: 0-13—088892-3.

[3] Jason Frand, University of California, Los Angles – Class Notes; http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm. (January 06)

[4] ABC News, “US Plans Massive Data Sweep”, 15th February 2006.

[5] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, “Crime Data Mining: An Overview and Case Studies.”; Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona. (March 2006)

[6] Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M., “Crime data mining: A general framework and some examples”; Volume 37, Issue 4, April 2004 Page(s):50 – 56.

[7] Brown, D.E, “The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals”; IEEE International Conference on Systems, Man, and Cybernetics; Volume 3, 11-14 Oct. 1998 Page(s):2848 - 2853 Vol.3.

[8] Chan, P.K.; Fan, W., Prodromidis A.L.; Stolfo S.J., “Distributed data mining in credit card fraud detection”; Intelligent Systems and Their Applications, IEEE; Volume 14, Issue 6, Nov.-Dec. 1999 Page(s):67 – 74.

[9] U.S. Congress, Office of Technology Assessment, “Information Technologies for Control of Money Laundering”; OTA-ITC-630 (Washington, DC: U.S. Government Printing Office, September 1995).

[10] Anil K. Jain, Robert P.W. Duin, Jianchang Mao, “Statistical Pattern Recognition: A Review”; IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, January 2000.

[11] Anil K. Jain, M. N. Murthy, P. J. Flynn, “Data Clustering: A Review”; ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[12] CGG Aitken, Dept. of Mathematics and Statistics, The University of Edinburg, UK EH9 3JZ; T Connolly, Boimathematics and Statistics Scotland, Scottish Corp Research Institue, UK DD2 5DA; A Gammerman, Dept. of Computer Science, Royal Holloway and Bedford New College, University of London, Uk, TW20 0EX; G Zang, Dept. of Electrical Engineering and Computer Science, LeHigh University, US 18015; D Bailey and R Gordon, Derbyshire Constabulary Headquaters, Uk DE5 3RS; R Oldfield, Police Research Group London, UK SW1H 9AT, "Science & Justice 1996”; 36(4): Page 245-255.

[13] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining".

[14] Kutner, Nachtsheim, Neterr, "Applied Linear Regression Models"; 4th Edition, McGraw-Hill, ISBN 0-07-301344-7.

[15] Caroline Clabaugh, Dave Myszewski, Jimmy Pang; "Presentation on Neural Networks", cse.stanford.edu/class/sophomore-college/projects-00/neural-networks/, CSE Dept. University of Stanford. (Project Year 2000)

[16] Gisele L. Pappa, Alex A. Freitas, "Towards a Genetic Programming Algorithm for Automatically Evolving Rule Induction Algorithms"; University of Kent, www.ke.informatik.tu-darmstadt.de/events/ECML-PKDD-04-WS/Proceedings/pappa.pdf. (March 2006)

[17] Corin R. Anderson, Pedro Domingos, Daniel S. Weld, "Relational Markov Models and their Application to Adaptive Web Navigation"; SIGKDD 2002 Edmonton, Alberta, Canada.

[18] Michael J. Hernandez, "Database Design for Mere Mortals"; 2nd Edition, Addison – Wesley, ISBN: 0-201-75284-0.

[19] Laurene Fausett, "Fundamentals of Neural Networks"; Prentice Hall, ISBN: 0-13-334186-0.

[20] Ian H. Witten and Eibe Frank (2005), "Data Mining: Practical machine learning tools and techniques"; 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[21] West Virginia State Police Department; http://www.wvstatepolice.com/wanted/wanted.shtml. (August 2006)

[22] Bureau of Justice Statistics, US; http://www.ojp.usdoj.gov/bjs/welcome.html. (July 2006)

[23] Dr. Kristof Van Laerhoven, Post-Doc, Computer Science, Darmstadt University of Technology, Germany; http://www.comp.lancs.ac.uk/~kristof/research/notes/index.html. (January 2006)

[24] Juha Vesanto, Esa Alhoniemi, "Clustering of the Self-Organizing Map." IEEE Transactiosn on Neural Networks, Vol. 11, No. 3, May 2000.

[25] The Queens University of Belfast UK – Data Mining Notes, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html. (February 2006)

[26] Dr. Kardi Tekonomo, http://people.revoledu.com/kardi/tutorial/kMean/index.html. (February 2006)

[27] Wikipedia.org, http://en.wikipedia.org/wiki/Main_Page. (August 2006)

APPENDIX A: Test Data from Query Table "tblQuery"

ACN	Year	Mth.	Dt.	Hr.	Min.	Day	County	OffType	Gen.	Race	Age	Crime Bkgd
1001	2004	1	2	12	28	Tue	Harrison	Assault	M	White	31	No
1002	2004	1	3	2	7	Sat	Monongalia	Drugs	M	Black	21	No
1003	2004	1	3	2	22	Wed	Cabell	Murder	M	Black	23	Yes
1004	2004	1	4	10	55	Thru	Monongalia	Drugs	F	AorPI	20	No
1005	2004	1	10	4	22	Sat	Marion	Burglary	M	White	24	Yes
1006	2004	1	10	23	38	Sun	Hancock	Murder	M	White	36	Yes
1007	2004	1	13	7	50	Tue	Berkley	Accident	F	AorPI	52	No
1008	2004	1	16	3	9	Fri	Monongalia	Sex_Assault	M	White	40	Yes
1009	2004	1	20	7	13	Tue	Mercer	Accident	M	White	52	No
1010	2004	1	25	11	16	Sun	Harrison	Arson	M	Black	22	Yes
1011	2004	1	25	20	34	Sun	Mercer	Weap_Threat	M	Black	27	Yes
1012	2004	1	27	9	59	Tue	Cabell	Assault	M	Black	29	Yes
1013	2004	1	27	3	29	Tue	Cabell	Burglary	M	Black	48	Yes
1014	2004	2	1	7	42	Sun	Marion	Murder	M	Black	43	Yes
1015	2004	2	2	1	1	Mon	Monongalia	Accident	M	AorPI	22	No
1016	2004	2	7	22	18	Sat	Monongalia	Sex_Assault	M	White	21	No
1017	2004	2	7	5	29	Sat	Monongalia	Rape	M	Black	21	Yes
1018	2004	2	9	18	5	Mon	Hancock	Accident	F	Black	24	No
1019	2004	2	10	4	29	Tue	Mercer	Burglary	M	Black	33	Yes
1020	2004	2	10	9	49	Tue	Cabell	Burglary	M	AorPI	25	No
1021	2004	2	11	10	53	Wed	Marion	Burglary	M	AorPI	56	No
1022	2004	2	17	13	20	Tue	Monongalia	Burglary	F	Black	37	Yes
1023	2004	2	21	3	5	Sat	Monongalia	Rape	M	White	46	Yes
1024	2004	2	23	19	51	Mon	Kanawha	Sex_Assault	M	Black	38	Yes
1025	2004	3	1	16	27	Mon	Mercer	Accident	M	White	28	No
1026	2004	3	13	1	10	Sat	Kanawha	Sex_Assault	M	Black	21	Yes
1027	2004	3	31	21	8	Wed	Cabell	Arson	M	White	39	Yes
1028	2004	4	4	0	27	Sun	Hancock	Weap_Threat	M	Black	23	Yes
1029	2004	4	6	1	57	Tue	Cabell	Burglary	M	White	41	No
1030	2004	4	9	1	32	Fri	Monongalia	Rape	M	Black	28	Yes
1031	2004	4	11	2	39	Sun	Mercer	Burglary	M	Hspnc	44	Yes
1032	2004	4	17	15	23	Sat	Monongalia	Drugs	M	Black	19	Yes
1033	2004	4	17	23	16	Sat	Cabell	Arson	M	Black	29	Yes
1034	2004	4	17	23	26	Sat	Raleigh	Rape	M	Black	43	Yes
1035	2004	4	20	1	13	Tue	Harrison	Arson	M	Black	32	Yes
1036	2004	4	29	22	20	Thru	Hancock	Drugs	M	Black	29	Yes
1037	2004	5	5	15	15	Wed	Monongalia	Burglary	M	Hspnc	36	Yes
1038	2004	5	8	22	52	Sat	Monongalia	Drugs	M	Black	30	Yes
1039	2004	5	16	1	10	Sun	Cabell	Burglary	M	Black	21	Yes
1040	2004	5	18	1	24	Tue	Harrison	Arson	M	Black	24	Yes

1041	2004	5	19	15	46	Wed	Mercer	Assault	M	Hspnc	37	No
1042	2004	5	21	1	50	Fri	Monongalia	Rape	M	Hspnc	46	Yes
1043	2004	5	27	20	46	Thru	Kanawha	Sex_Assault	M	Hspnc	33	Yes
1044	2004	5	27	23	14	Thru	Monongalia	Rape	M	White	23	Yes
1045	2004	6	2	7	36	Wed	Kanawha	Accident	F	White	36	Yes
1046	2004	6	2	7	53	Wed	Kanawha	Accident	F	AorPI	41	No
1047	2004	6	2	7	18	Wed	Cabell	Accident	M	White	57	No
1048	2004	6	2	23	50	Wed	Cabell	Burglary	M	Hspnc	26	Yes
1049	2004	6	5	15	52	Sat	Hancock	Assault	M	Hspnc	37	Yes
1050	2004	6	8	8	44	Tue	Berkley	Weap_Threat	M	Hspnc	59	Yes
1051	2004	6	8	9	0	Tue	Cabell	Burglary	F	Hspnc	29	Yes
1052	2004	6	8	19	3	Tue	Monongalia	Accident	M	Black	29	No
1053	2004	6	11	21	38	Fri	Monongalia	Sex_Assault	M	Black	23	Yes
1054	2004	6	11	2	31	Fri	Monongalia	Rape	M	Black	23	Yes
1055	2004	6	12	1	54	Sat	Monongalia	Sex_Assault	M	White	46	Yes
1056	2004	6	15	8	11	Tue	Monongalia	Accident	M	AorPI	24	Yes
1057	2004	6	16	19	44	Wed	Hancock	Assault	M	Hspnc	30	Yes
1058	2004	6	18	21	34	Fri	Hancock	Weap_Threat	M	White	49	Yes
1059	2004	6	24	23	15	Thru	Monongalia	Drugs	M	Hspnc	21	Yes
1060	2004	6	24	22	29	Thru	Monongalia	Drugs	M	AorPI	22	No
1061	2004	6	25	23	8	Fri	Monongalia	Sex_Assault	M	White	47	Yes
1062	2004	6	30	7	15	Wed	Mercer	Accident	M	Hspnc	52	No
1063	2004	6	30	14	46	Wed	Marion	Burglary	F	Hspnc	34	Yes
1064	2004	7	3	20	51	Sat	Kanawha	Rape	M	Black	39	Yes
1065	2004	7	4	23	27	Sun	Cabell	Burglary	M	Hspnc	36	Yes
1066	2004	7	5	9	38	Mon	Raleigh	Accident	M	White	31	Yes
1067	2004	7	6	19	38	Tue	Mercer	Accident	M	White	20	No
1068	2004	7	10	21	16	Sat	Monongalia	Sex_Assault	M	Black	37	Yes
1069	2004	7	11	2	21	Sun	Cabell	Burglary	F	White	31	Yes
1070	2004	7	26	0	46	Mon	Harrison	Accident	M	Black	37	No
1071	2004	7	31	1	39	Sat	Kanawha	Sex_Assault	M	Hspnc	46	Yes
1072	2004	8	1	22	50	Sun	Berkley	Arson	M	White	61	Yes
1073	2004	8	7	22	42	Sat	Monongalia	Rape	M	Hspnc	37	Yes
1074	2004	8	8	15	43	Sun	Marion	Arson	M	White	24	Yes
1075	2004	8	9	23	38	Mon	Marion	Accident	M	Black	18	No
1076	2004	8	11	3	51	Wed	Berkley	Burglary	F	White	29	Yes
1077	2004	8	31	3	11	Tue	Cabell	Burglary	M	AorPI	21	Yes
1078	2004	9	8	7	56	Wed	Monongalia	Accident	M	White	19	No
1079	2004	9	11	2	19	Sat	Monongalia	Rape	M	Black	38	Yes
1080	2004	9	14	22	49	Tue	Berkley	Drugs	M	Black	25	Yes
1081	2004	9	17	19	39	Fri	Monongalia	Sex_Assault	M	Black	27	Yes
1082	2004	9	18	23	10	Sat	Monongalia	Accident	M	White	67	Yes
1083	2004	9	21	2	59	Tue	Marion	Murder	F	Black	34	Yes
1084	2004	9	21	20	23	Tue	Cabell	Burglary	M	Black	24	Yes
1085	2004	9	24	19	19	Fri	Kanawha	Drugs	M	AorPI	33	Yes
1086	2004	9	26	2	48	Sun	Harrison	Accident	M	Black	15	Yes
1087	2004	10	3	0	7	Sun	Berkley	Accident	M	White	49	No

1088	2004	10	3	23	27	Sun	Monongalia	Assault	M	Hspnc	33	Yes
1089	2004	10	17	4	1	Sun	Cabell	Burglary	M	White	44	No
1090	2004	10	17	15	49	Sun	Kanawha	Weap_Threat	M	Black	20	Yes
1091	2004	10	18	8	36	Mon	Hancock	Accident	M	White	28	Yes
1092	2004	10	26	3	46	Tue	Cabell	Burglary	M	Black	21	No
1093	2004	10	29	2	18	Fri	Marion	Sex_Assault	M	Black	27	Yes
1094	2004	10	30	2	7	Sat	Hancock	Drugs	M	Black	22	Yes
1095	2004	11	2	21	2	Tue	Berkley	Weap_Threat	M	White	54	Yes
1096	2004	11	3	16	6	Wed	Kanawha	Arson	M	Black	34	No
1097	2004	11	15	18	54	Tue	Cabell	Burglary	M	AorPI	53	Yes
1098	2004	11	27	2	9	Sat	Mercer	Rape	M	White	28	Yes
1099	2004	11	28	1	30	Sun	Raleigh	Accident	M	White	27	No
1100	2004	11	30	3	56	Tue	Mercer	Burglary	M	White	24	No
1101	2004	12	5	23	56	Mon	Marion	Murder	M	Black	38	Yes
1102	2004	12	6	18	56	Tue	Hancock	Burglary	M	White	17	No
1103	2004	12	23	18	43	Thru	Monongalia	Drugs	M	White	21	Yes
1104	2004	12	28	9	1	Tue	Marion	Assault	M	White	32	Yes
1105	2004	12	28	1	46	Tue	Harrison	Arson	M	White	16	Yes
1106	2004	12	30	12	6	Thru	Hancock	Weap_Threat	M	Black	28	Yes
1107	2004	12	30	10	34	Thru	Hancock	Drugs	M	White	62	Yes
1108	2004	12	30	2	7	Thru	Marion	Drugs	M	Black	30	Yes
1109	2004	12	31	18	19	Fri	Raleigh	Assault	M	Hspnc	22	No
1110	2004	12	31	14	43	Fri	Monongalia	Sex_Assault	M	Black	30	Yes
1111	2005	1	2	0	32	Sun	Hancock	Rape	M	Black	38	Yes
1112	2005	1	2	14	46	Sun	Harrison	Arson	M	White	25	No
1113	2005	1	4	15	23	Tue	Kanawha	Weap_Threat	M	Black	27	Yes
1114	2005	1	10	7	51	Mon	Harrison	Accident	M	White	67	No
1115	2005	1	15	16	59	Sat	Cabell	Assault	M	Black	36	Yes
1116	2005	1	16	3	36	Sun	Cabell	Burglary	M	Hspnc	53	Yes
1117	2005	1	21	1	43	Fri	Monongalia	Rape	M	Black	29	Yes
1118	2005	1	25	14	58	Tue	Hancock	Weap_Threat	M	Hspnc	49	Yes
1119	2005	1	25	19	24	Tue	Mercer	Sex_Assault	M	Black	33	Yes
1120	2005	1	30	2	30	Sun	Monongalia	Arson	M	White	23	Yes
1121	2005	1	31	20	17	Mon	Hancock	Accident	F	Black	20	No
1122	2005	2	1	4	40	Tue	Berkley	Burglary	M	Hspnc	35	Yes
1123	2005	2	2	18	7	Wed	Monongalia	Sex_Assault	M	Black	31	Yes
1124	2005	2	7	17	1	Mon	Harrison	Accident	M	Black	47	Yes
1125	2005	2	7	4	5	Mon	Mercer	Burglary	M	Black	32	Yes
1126	2005	2	9	15	51	Wed	Kanawha	Accident	M	White	54	No
1127	2005	2	10	18	41	Thru	Hancock	Drugs	M	Black	44	Yes
1128	2005	2	11	7	22	Fri	Raleigh	Rape	M	Black	28	Yes
1129	2005	2	12	19	19	Sat	Monongalia	Drugs	F	Black	22	Yes
1130	2005	2	15	12	33	Tue	Berkley	Sex_Assault	M	Black	27	No
1131	2005	2	19	22	22	Sat	Raleigh	Drugs	M	AorPI	24	Yes
1132	2005	2	26	0	49	Sat	Monongalia	Sex_Assault	M	White	50	Yes
1133	2005	3	1	22	53	Tue	Marion	Murder	M	Black	57	Yes
1134	2005	3	13	22	46	Sun	Harrison	Murder	M	White	51	Yes

1135	2005	3	13	9	35	Sun	Kanawha	Assault	M	White	42	Yes
1136	2005	3	29	2	52	Tue	Monongalia	Burglary	M	Hspnc	34	No
1137	2005	3	31	23	57	Thru	Kanawha	Rape	M	White	52	Yes
1138	2005	4	8	20	8	Fri	Cabell	Drugs	F	White	22	No
1139	2005	4	9	0	59	Sat	Monongalia	Sex_Assault	M	Black	39	Yes
1140	2005	4	12	12	25	Tue	Raleigh	Burglary	M	Black	39	Yes
1141	2005	4	17	0	46	Sun	Harrison	Accident	M	Black	19	Yes
1142	2005	4	18	18	51	Mon	Hancock	Weap_Threat	M	White	53	Yes
1143	2005	4	23	20	25	Sat	Raleigh	Sex_Assault	M	Black	45	Yes
1144	2005	4	26	17	14	Tue	Kanawha	Accident	M	Black	25	No
1145	2005	5	5	23	56	Thru	Hancock	Drugs	F	White	21	Yes
1146	2005	5	16	9	30	Mon	Berkley	Assault	M	Black	24	Yes
1147	2005	5	18	8	43	Wed	Cabell	Accident	M	Black	33	Yes
1148	2005	5	19	23	37	Thru	Kanawha	Rape	M	Black	37	Yes
1149	2005	5	21	4	22	Sat	Cabell	Burglary	M	Black	22	Yes
1150	2005	5	21	1	54	Sat	Marion	Murder	M	White	46	Yes
1151	2005	5	25	18	44	Wed	Mercer	Burglary	M	Black	24	Yes
1152	2005	5	27	0	2	Fri	Raleigh	Rape	M	White	38	Yes
1153	2005	5	27	20	19	Fri	Mercer	Drugs	M	AorPI	21	No
1154	2005	6	4	23	32	Sat	Hancock	Accident	M	Hspnc	39	Yes
1155	2005	6	7	7	10	Tue	Marion	Accident	M	Black	30	No
1156	2005	6	7	19	31	Tue	Monongalia	Sex_Assault	M	Black	41	Yes
1157	2005	6	8	9	21	Wed	Cabell	Accident	M	Black	36	Yes
1158	2005	6	9	17	38	Thru	Kanawha	Drugs	M	White	23	Yes
1159	2005	6	11	7	3	Sat	Kanawha	Sex_Assault	M	Black	39	Yes
1160	2005	6	11	1	22	Sat	Raleigh	Sex_Assault	M	Hspnc	44	Yes
1161	2005	6	15	11	11	Wed	Harrison	Burglary	M	Black	51	Yes
1162	2005	6	16	19	34	Thru	Harrison	Drugs	M	Black	24	Yes
1163	2005	6	17	23	45	Fri	Berkley	Assault	M	Black	37	Yes
1164	2005	6	21	12	7	Tue	Monongalia	Rape	M	Black	24	Yes
1165	2005	6	24	22	12	Fri	Kanawha	Arson	M	Black	32	Yes
1166	2005	6	26	8	27	Sun	Raleigh	Sex_Assault	M	Hspnc	52	Yes
1167	2005	6	26	15	58	Sun	Mercer	Burglary	M	Hspnc	44	Yes
1168	2005	7	2	0	53	Sat	Hancock	Rape	M	Hspnc	52	Yes
1169	2005	7	3	11	9	Sun	Hancock	Assault	M	White	22	Yes
1170	2005	7	4	17	56	Mon	Cabell	Accident	M	White	19	No
1171	2005	7	5	23	23	Tue	Marion	Murder	M	White	39	Yes
1172	2005	7	6	2	25	Wed	Harrison	Burglary	M	Hspnc	53	Yes
1173	2005	7	10	10	38	Sun	Cabell	Assault	M	Black	25	Yes
1174	2005	7	11	8	28	Mon	Raleigh	Accident	F	AorPI	63	No
1175	2005	7	24	10	10	Sun	Harrison	Assault	M	Black	34	Yes
1176	2005	7	26	17	16	Tue	Berkley	Accident	M	Black	33	No
1177	2005	7	31	12	0	Sun	Harrison	Arson	M	Black	45	Yes
1178	2005	8	1	1	17	Mon	Monongalia	Drugs	M	Black	19	No
1179	2005	8	7	13	29	Sun	Hancock	Weap_Threat	F	Black	33	Yes
1180	2005	8	8	0	7	Mon	Kanawha	Sex_Assault	M	Black	44	Yes
1181	2005	8	9	15	23	Tue	Berkley	Assault	M	Black	20	No

1182	2005	8	11	21	23	Thru	Raleigh	Drugs	M	AorPI	28	Yes
1183	2005	8	23	11	41	Tue	Cabell	Burglary	M	White	41	Yes
1184	2005	8	24	16	16	Wed	Marion	Accident	M	White	47	No
1185	2005	8	27	1	47	Sat	Monongalia	Rape	M	Black	32	Yes
1186	2005	9	3	0	15	Sat	Monongalia	Drugs	M	Black	16	Yes
1187	2005	9	4	6	17	Sun	Harrison	Sex_Assault	M	Black	34	Yes
1188	2005	9	11	19	48	Sun	Berkley	Assault	M	Black	29	Yes
1189	2005	9	14	22	8	Wed	Berkley	Sex_Assault	M	Hspnc	44	Yes
1190	2005	9	17	14	38	Sat	Mercer	Accident	F	White	28	No
1191	2005	9	23	0	34	Fri	Kanawha	Drugs	F	Black	27	Yes
1192	2005	9	28	4	8	Wed	Mercer	Burglary	M	Black	24	Yes
1193	2005	9	28	3	30	Wed	Cabell	Weap_Threat	M	Black	50	Yes
1194	2005	9	30	17	52	Fri	Mercer	Arson	M	Hspnc	57	Yes
1195	2005	10	3	20	54	Mon	Hancock	Accident	M	White	51	No
1196	2005	10	14	0	25	Fri	Monongalia	Rape	M	Hspnc	52	Yes
1197	2005	10	15	23	54	Sat	Hancock	Assault	M	White	24	Yes
1198	2005	10	22	2	16	Sat	Cabell	Drugs	M	Black	20	Yes
1199	2005	10	22	22	45	Sat	Kanawha	Sex_Assault	M	White	57	Yes
1200	2005	10	23	0	6	Sun	Marion	Assault	M	Black	24	No
1201	2005	10	27	23	26	Thru	Monongalia	Drugs	M	Hspnc	21	Yes
1202	2005	10	27	20	35	Thru	Hancock	Drugs	M	AorPI	29	Yes
1203	2005	10	30	3	28	Sun	Monongalia	Murder	M	Black	41	Yes
1204	2005	11	3	23	49	Thru	Raleigh	Drugs	M	White	21	No
1205	2005	11	27	3	3	Sun	Berkley	Rape	M	Black	47	Yes
1206	2005	11	29	1	55	Tue	Kanawha	Accident	M	Black	60	Yes
1207	2005	11	30	3	19	Wed	Cabell	Burglary	M	Black	22	Yes
1208	2005	11	30	15	0	Wed	Hancock	Arson	M	Black	41	Yes
1209	2005	11	30	23	1	Wed	Berkley	Drugs	M	Black	28	Yes
1210	2005	12	3	21	33	Sat	Monongalia	Drugs	M	Black	23	Yes
1211	2005	12	17	0	40	Sat	Monongalia	Drugs	F	Black	15	No
1212	2005	12	22	19	27	Thru	Monongalia	Drugs	F	White	23	Yes
1213	2005	12	22	14	15	Thru	Mercer	Rape	M	Black	27	Yes
1214	2005	12	26	6	26	Mon	Berkley	Accident	M	Black	67	No
1215	2005	12	26	7	47	Mon	Harrison	Accident	M	White	63	No
1216	2005	12	30	0	55	Fri	Monongalia	Drugs	M	AorPI	23	No
1217	2005	12	31	1	29	Sat	Kanawha	Rape	M	White	45	Yes
1218	2005	12	31	22	12	Sat	Berkley	Sex_Assault	M	Black	49	Yes
1219	2005	12	31	23	53	Sat	Kanawha	Rape	M	Black	27	No
1220	2006	1	1	14	41	Sun	Mercer	Weap_Threat	M	Black	24	Yes
1221	2006	1	1	23	26	Sun	Marion	Burglary	M	AorPI	50	Yes
1222	2006	1	3	15	2	Tue	Monongalia	Drugs	M	Black	18	Yes
1223	2006	1	7	0	19	Sat	Cabell	Arson	M	White	51	No
1224	2006	1	7	21	27	Sat	Monongalia	Rape	M	Black	53	Yes
1225	2006	1	12	19	50	Thru	Berkley	Drugs	M	Black	26	Yes
1226	2006	1	14	20	56	Sat	Harrison	Assault	M	White	33	Yes
1227	2006	1	17	0	20	Tue	Mercer	Weap_Threat	M	Black	22	Yes
1228	2006	1	20	3	34	Fri	Hancock	Weap_Threat	M	White	56	Yes

1229	2006	1	20	3	47	Fri	Monongalia	Drugs	M	Black	19	No
1230	2006	1	30	6	40	Mon	Harrison	Accident	M	Hspnc	19	No
1231	2006	1	31	18	14	Tue	Marion	Burglary	M	Hspnc	53	Yes
1232	2006	2	4	0	53	Sat	Monongalia	Rape	M	Black	45	Yes
1233	2006	2	9	18	38	Thru	Kanawha	Drugs	M	White	24	Yes
1234	2006	2	11	23	59	Sat	Raleigh	Rape	M	Black	29	Yes
1235	2006	2	12	23	8	Sun	Monongalia	Sex_Assault	M	AorPI	36	Yes
1236	2006	2	16	0	27	Thru	Harrison	Drugs	M	Black	16	No
1237	2006	2	23	19	57	Thru	Mercer	Accident	M	Hspnc	56	No
1238	2006	2	23	21	37	Thru	Marion	Weap_Threat	M	Black	27	Yes
1239	2006	2	28	2	45	Tue	Cabell	Burglary	M	AorPI	46	Yes
1240	2006	3	2	21	42	Thru	Hancock	Rape	M	Black	36	Yes
1241	2006	3	4	22	39	Sat	Monongalia	Rape	M	Black	49	Yes
1242	2006	3	5	14	17	Sun	Monongalia	Assault	M	Black	33	No
1243	2006	3	6	9	19	Mon	Marion	Accident	M	AorPI	39	No
1244	2006	3	6	20	9	Mon	Raleigh	Accident	M	Black	49	No
1245	2006	3	16	23	4	Thru	Monongalia	Drugs	M	AorPI	23	No
1246	2006	3	29	14	56	Wed	Hancock	Burglary	F	White	22	Yes
1247	2006	3	30	22	34	Thru	Marion	Drugs	M	Black	15	No
1248	2006	4	1	18	53	Sat	Mercer	Sex_Assault	M	White	45	Yes
1249	2006	4	1	19	23	Sat	Monongalia	Rape	M	Black	38	Yes
1250	2006	4	2	2	38	Sun	Marion	Murder	M	Hspnc	49	Yes
1251	2006	4	2	2	46	Sun	Monongalia	Burglary	M	White	15	No
1252	2006	4	15	1	43	Sat	Mercer	Rape	M	Hspnc	44	Yes
1253	2006	4	18	7	15	Tue	Raleigh	Accident	M	Black	20	Yes
1254	2006	4	21	1	17	Fri	Marion	Drugs	M	Hspnc	28	Yes
1255	2006	4	26	8	59	Wed	Cabell	Accident	M	White	41	Yes
1256	2006	5	1	20	16	Mon	Monongalia	Sex_Assault	M	Black	31	Yes
1257	2006	5	6	1	52	Sat	Monongalia	Drugs	M	Black	53	Yes
1258	2006	5	8	0	4	Mon	Mercer	Accident	M	Hspnc	42	No
1259	2006	5	8	11	49	Mon	Cabell	Burglary	M	White	24	Yes
1260	2006	5	9	16	46	Tue	Harrison	Accident	F	AorPI	44	No
1261	2006	5	12	19	15	Fri	Mercer	Arson	M	Hspnc	28	Yes
1262	2006	5	12	4	24	Fri	Monongalia	Sex_Assault	M	White	27	Yes
1263	2006	5	14	0	23	Sun	Cabell	Weap_Threat	F	Black	31	No
1264	2006	5	14	14	39	Sun	Kanawha	Assault	M	Hspnc	24	Yes
1265	2006	5	15	15	19	Mon	Marion	Accident	M	Black	50	No
1266	2006	5	15	15	38	Fri	Raleigh	Rape	M	Black	48	Yes
1267	2006	5	19	22	11	Fri	Berkley	Accident	M	White	51	No
1268	2006	5	23	12	24	Tue	Harrison	Arson	M	Black	23	No
1269	2006	5	23	18	5	Tue	Cabell	Burglary	M	White	45	Yes
1270	2006	5	26	2	56	Fri	Monongalia	Drugs	M	White	23	Yes
1271	2006	5	27	15	15	Sat	Raleigh	Rape	M	Black	50	Yes
1272	2006	5	29	22	6	Mon	Marion	Murder	M	Hspnc	57	Yes
1273	2006	6	1	19	2	Thru	Monongalia	Drugs	M	AorPI	21	Yes
1274	2006	6	3	19	3	Sat	Monongalia	Rape	M	Black	45	No
1275	2006	6	9	0	25	Fri	Berkley	Drugs	M	Black	14	No

1276	2006	6	12	21	57	Mon	Cabell	Rape	M	Black	54	Yes
1277	2006	6	17	0	45	Sat	Mercer	Sex_Assault	M	Black	38	Yes
1278	2006	6	17	1	33	Sat	Harrison	Rape	M	Black	37	Yes
1279	2006	6	23	18	8	Fri	Berkley	Assault	M	Black	24	Yes
1280	2006	6	25	9	44	Sun	Monongalia	Drugs	M	AorPI	22	Yes
1281	2006	6	28	9	36	Wed	Raleigh	Accident	M	Hspnc	57	No
1282	2006	6	28	20	14	Wed	Kanawha	Assault	M	Hspnc	51	Yes
1283	2006	7	1	18	14	Sat	Monongalia	Burglary	M	Hspnc	62	Yes
1284	2006	7	2	10	42	Sun	Hancock	Weap_Threat	M	White	51	Yes
1285	2006	7	5	10	27	Wed	Monongalia	Accident	F	AorPI	24	No
1286	2006	7	5	19	27	Wed	Mercer	Sex_Assault	M	White	26	Yes
1287	2006	7	8	0	39	Sat	Berkley	Rape	M	Black	39	Yes
1288	2006	7	11	14	23	Tue	Hancock	Accident	F	Hspnc	59	Yes
1289	2005	7	16	23	35	Sat	Monongalia	Sex_Assault	M	Black	36	No
1290	2006	7	24	0	3	Mon	Harrison	Accident	M	Black	53	No
1291	2006	7	30	20	56	Sun	Mercer	Weap_Threat	M	Black	25	Yes
1292	2006	8	2	1	56	Wed	Cabell	Burglary	M	Hspnc	34	Yes
1293	2006	8	6	21	6	Sun	Harrison	Accident	M	Black	29	No
1294	2006	8	6	0	33	Sun	Berkley	Accident	M	White	24	Yes
1295	2006	8	10	23	5	Thru	Marion	Drugs	M	Black	33	Yes
1296	2006	8	11	21	7	Fri	Monongalia	Rape	M	White	56	Yes
1297	2006	8	25	22	23	Fri	Cabell	Sex_Assault	M	Hspnc	27	Yes
1298	2006	8	26	9	56	Sat	Monongalia	Rape	M	Black	41	No
1299	2006	8	28	19	52	Mon	Mercer	Accident	M	White	46	No
1300	2006	9	3	1	19	Sun	Harrison	Burglary	M	Hspnc	22	Yes
1301	2006	9	3	22	15	Sun	Marion	Murder	M	Black	39	Yes
1302	2006	9	6	11	33	Wed	Harrison	Accident	M	White	19	Yes
1303	2006	9	8	10	40	Fri	Hancock	Weap_Threat	F	White	33	Yes
1304	2006	9	23	10	52	Sat	Monongalia	Sex_Assault	M	Black	43	Yes
1305	2006	9	24	0	5	Sun	Monongalia	Arson	M	White	19	No
1306	2006	9	25	9	49	Mon	Mercer	Accident	M	Black	45	No
1307	2006	9	29	2	0	Fri	Hancock	Drugs	M	Black	14	No
1308	2006	10	5	23	23	Thru	Monongalia	Sex_Assault	M	Black	33	Yes
1309	2006	10	15	22	54	Sun	Cabell	Accident	M	Black	44	Yes
1310	2006	10	17	2	38	Tue	Mercer	Burglary	M	Black	20	Yes
1311	2006	10	21	0	10	Sat	Raleigh	Rape	M	Black	42	Yes
1312	2006	10	23	1	48	Mon	Cabell	Accident	M	White	51	No
1313	2006	10	24	13	6	Tue	Harrison	Accident	M	Black	62	Yes
1314	2006	10	30	13	17	Mon	Cabell	Burglary	M	White	31	Yes
1315	2006	11	2	18	23	Thru	Monongalia	Drugs	M	AorPI	21	No
1316	2006	11	6	19	34	Mon	Kanawha	Sex_Assault	M	Hspnc	52	Yes
1317	2006	11	21	21	54	Tue	Hancock	Weap_Threat	M	Black	30	Yes
1318	2006	11	21	15	16	Tue	Marion	Burglary	M	AorPI	22	Yes
1319	2006	11	23	0	59	Fri	Raleigh	Rape	M	Black	50	Yes
1320	2006	11	30	21	35	Thru	Monongalia	Drugs	M	AorPI	24	Yes
1321	2006	12	6	13	27	Wed	Harrison	Accident	M	Black	15	No
1322	2006	12	23	19	40	Sat	Hancock	Drugs	M	Hspnc	53	Yes

1323	2006	12	23	19	10	Sat	Monongalia	Sex_Assault	M	Black	45	Yes
1324	2006	12	25	9	12	Mon	Berkley	Accident	M	Black	25	No
1325	2006	12	25	11	18	Mon	Monongalia	Weap_Threat	M	Black	29	Yes
1326	2006	12	30	11	56	Sat	Monongalia	Rape	M	Black	43	No
1327	2006	12	30	22	45	Sat	Cabell	Accident	M	White	24	No
1328	2006	12	31	8	54	Sun	Mercer	Sex_Assault	M	Black	33	Yes
1329	2006	12	31	0	22	Sun	Harrison	Accident	M	White	56	Yes
1330	2006	12	31	0	53	Sun	Raleigh	Rape	M	Black	37	Yes

APPENDIX B: Requirements Document

This is a customized and brief form of a requirements document; the original documents are much more extensive and cover other modules such as Database Requirement, Navigation Requirement, etc. Ones document encapsulates the all the aspects of the model into one single document. This helps in clear understanding of various functions and constrains. Finally, the document is an outline to what the model is all about.

Objective

The objective of the model is to allow the end user to navigate through all the data searching and mining technique, use them with the data stored in the pre-selected data table and store the results of his / her choice in order to refer back to it at the end.

Scope

a. Allow the user to select the data mining and search technique of his choice.
b. Provide a data source where the user is able to execute the selected algorithm without any concern of the data, algorithms or techniques.
c. Provide ability of storing the results for further reference.

End-users

This model is intended for use by detectives of different crime investigation agencies such as State Police Dept. and FBI.

Assumptions

The user should have basic knowledge of Searching / Querying techniques and ability of interpreting the results of Data Mining algorithms.

Constraints

The tool works only with predefined data table. Web functionality cannot be provided to all the features in the model.

Model Requirements

a. The tool should have an appropriate navigational structure.
b. The tool should have a database capable of addressing issues as per the objective.
c. The different algorithms used in the tool should generate correct results.
d. The obtained results should serve the objective.

Model Functionality

The tool design would be such that the end user must have basic data querying and interpretation knowledge. Initially, the Data Mining methods would assist to identify relationship between different variables of the Query Table and find out any non-obvious crime patterns in it. The searching algorithms would then be used by putting in specific queries that based on the patterns would find specific details about the pattern.

Testing Considerations

a. Does the tool help the end user meet the functional requirements?
b. Is the user interface consistent and user friendly?
c. Is the tool database robust so as to meet different or specific needs of the end user?
d. Are the results provided by the tool correct and serve purpose of the end user?

APPENDIX C: Association Rules Output

Association Rule output with columns “OffDay”, “OffType”, “County” and “Race”.

Monongalia – Black	: 42
Sat – Black	: 41
Tue – Black	: 30
Sat – Monongalia	: 30
Sun – Black	: 27
Accident – White	: 26
Drugs – Black	: 26
Accident – Black	: 25
Monongalia – Drugs	: 24
Cabell – Burglary	: 22
Tue – Burglary	: 21

Association Rule output with columns “OffDay”, “OffType” and “County”.

Sat – Monongalia	: 30
Monongalia – Drugs	: 24
Cabell – Burglary	: 22
Tue – Burglary	: 21
Tue – Accident	: 13
Tue – Cabell	: 12
Sat – Drugs	: 12
Sun – Burglary	: 10
Sun – Cabell	: 8
Sun – Monongalia	: 8
Cabell – Accident	: 8

Association Rule output with columns “OffDay” and “OffType”.

Tue – Burglary	: 21
Tue – Accident	: 13
Sat – Drugs	: 12
Sun – Burglary	: 10
Sun – Accident	: 8
Sat – Accident	: 4
Sat – Burglary	: 3
Tue – Drugs	: 2
Sun – Drugs	: 1

Association Rule output with columns “OffDay” and “County”.

Sat – Monongalia	: 30
Tue – Cabell	: 12
Sun – Monongalia	: 8
Sun – Cabell	: 8
Tue – Monongalia	: 7
Sat – Cabell	: 6
Sat – Hancock	: 6
Sun – Hancock	: 6
Tue – Hancock	: 4

APPENDIX D: Decision Rule Output

Decision Rule Output with Field as “OffType” and Field Value as “Accident”.

Field Name: OffType Field Value: Accident

Age
> 35: 37; between 24 and 35: 17; between 14 and 23: 12;

County
Harrison: 14; Mercer: 9; Cabell: 8;

CriminalBkgd
Yes: 20; No: 46;

Gender
Male: 56; Female: 10;

OffDate
between 6 and 10: 18; between 26 and 31: 16; between 1 and 5: 11;

OffDay
Mon: 25; Wed: 14; Tue: 13;

OffHour
between 6 and 9: 18; between 15 and 18: 9; between 9 and 12: 8;

OffMonth
Jun: 10; Jul: 9; Sep: 6;

Race
White: 26; Black: 25; AorPI: 8;

Decision Rule Output with Field as “OffType” and Field Value as “Arson”.

Field Name: OffType Field Value: Arson

Age
between 24 and 35: 8; > 35: 6; between 14 and 23: 5;

County
Harrison: 7; Cabell: 3; Kanawha: 2;

CriminalBkgd
Yes: 14; No: 5;

Gender
Male: 19;

OffDate
between 26 and 31: 6; between 21 and 25: 4; between 1 and 5: 3;

OffDay
Sun:7; Tue:4; Fri:3;

OffHour
between 15 and 18: 4; between 21 and 24: 4; between 12 and 15: 3;

OffMonth
Jan: 4; May: 3; Apr: 2;

Race
Black: 9; White: 8; Hspnc: 2;

Decision Rule Output with Field as “OffType” and Field Value as “Assault”.

Field Name: OffType Field Value: Assault

Age
between 24 and 35: 15; > 35: 6; between 14 and 23: 3;

County
Berkley: 5; Hancock: 4; Kanawha: 3;

CriminalBkgd
Yes: 18; No: 6;

Gender
Male: 24;

OffDate
between 11 and 15: 6; between 1 and 5: 5; between 16 and 20: 4;

OffDay
Sun: 9; Tue: 4; Sat: 4;

OffHour
between 9 and 12: 7; between 18 and 21: 6; between 15 and 18: 4;

OffMonth
Jun: 5; Jan: 4; May: 3;

Race
Black: 11; Hspnc: 7; White: 6;

Decision Rule Output with Field as “OffType” and Field Value as “Burglary”.

Field Name: OffType Field Value: Burglary

Age
> 35: 20; between 24 and 35: 18; between 14 and 23: 11;

County
Cabell: 22; Mercer: 8; Marion: 6;

CriminalBkgd
Yes: 40; No: 9;

Gender
Male: 43; Female: 6;

OffDate
between 26 and 31: 13; between 1 and 5: 9; between 6 and 10: 7;

OffDay
Tue: 21; Wed: 12; Sun: 10;

OffHour
between 3 and 6: 14; between 12 and 15: 7; between 18 and 21: 6;

OffMonth
Feb: 7; May: 6; Jan: 5;

Race
Hspnc: 15; Black: 13; White: 13;

Decision Rule Output with Field as “OffType” and Field Value as “Drugs”.

Field Name: OffType Field Value: Drugs

Age
between 14 and 23: 32; between 24 and 35: 16; > 35: 4;

County
Monongalia: 24; Hancock: 8; Kanawha: 4;

CriminalBkgd
Yes: 36; No: 16;

Gender
Male: 45; Female: 7;

OffDate
between 26 and 31: 13; between 21 and 25: 10; between 1 and 5: 10;

OffDay
Sat: 12; Fri: 10; Tue: 2;

OffHour
between 21 and 24: 16; between 18 and 21: 14; between 9 and 12: 3;

OffMonth
Dec: 8; Jun: 7; May: 5;

Race
Black: 26; AorPI: 13; White: 9;

Decision Rule Output with Field as “OffType” and Field Value as “Murder”.

Field Name: OffType Field Value: Murder

Age
> 35: 11; between 24 and 35: 1; between 14 and 23: 1;

County
Marion: 9; Hancock: 1; Harrison: 1;

CriminalBkgd
Yes: 13;

Gender
Male: 12; Female: 1;

OffDate
between 1 and 5: 7; between 21 and 25: 2; between 26 and 31: 2;

OffDay
Sun: 6; Tue: 3; Mon: 2;

OffHour
between 21 and 24: 7; between 6 and 9: 1; between 3 and 6: 1;

OffMonth
Jan: 2; Mar: 2; May: 2;

Race
Black: 7; White: 4; Hspnc: 2;

Decision Rule Output with Field as “OffType” and Field Value as “Rape”.

Field Name: OffType Field Value: Rape

Age
> 35: 22; between 24 and 35: 9; between 14 and 23: 3;

County
Monongalia: 20; Raleigh: 9; Kanawha: 5;

CriminalBkgd
Yes: 40; No: 4;

Gender
Male: 44;

OffDate
between 26 and 31: 12; between 11 and 15: 9; between 21 and 25: 8;

OffDay
Sat: 24; Fri: 10; Sun: 3;

OffHour
between 21 and 24: 12; between 18 and 21: 3; between 3 and 6: 3;

OffMonth
May: 6; Feb: 5; Jun: 5;

Race
Black: 32; White: 7; Hspnc: 5;

Decision Rule Output with Field as “OffType” and Field Value as “Sex_Assault”.

Field Name: OffType Field Value: Sex_Assault

Age
> 35: 24; between 24 and 35: 14; between 14 and 23: 3;

County
Monongalia: 19; Kanawha: 8; Mercer: 5;

CriminalBkgd
Yes: 38; No: 2;

Gender
Male: 41;

OffDate
between 11 and 15: 9; between 21 and 25: 8; between 26 and 31: 8;

OffDay
Sat: 17; Fri: 8; Mon: 4;

OffHour
between 18 and 21: 12; between 21 and 24: 9; between 6 and 9: 4;

OffMonth
Jun: 8; Feb: 6; Jul: 4;

Race
Black: 24; White: 9; Hspnc: 7;

Decision Rule Output with Field as “OffType” and Field Value as “Weap_Threat”.

Field Name: OffType Field Value: Weap_Threat

Age
between 24 and 35: 11; > 35: 8; between 14 and 23: 3;

County
Hancock: 10; Mercer: 4; Berkley: 2;

CriminalBkgd
Yes: 20; No: 1

6]Gender
Male: 19; Female: 3;

OffDate
between 1 and 5: 5; between 16 and 20: 5; between 21 and 25: 5;

OffDay
Sun: 8; Tue: 6; Fri: 3;

OffHour
between 12 and 15: 4; between 21 and 24: 4; between 18 and 21: 3;

OffMonth
Jan: 6; Apr: 2; Jun: 2;

Race
Black: 14; White: 8; Hspnc: 2;

APPENDIX E: WEKA Output

Apriori Output with 6 attributes.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3,9
Instances:    330
Attributes:   6
              OffDay
              OffCounty
              OffType
              Gender
              Race
              CriminalBkgd
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (33 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 22

Size of set of large itemsets L(2): 32

Size of set of large itemsets L(3): 15

Size of set of large itemsets L(4): 1

Best rules found:

 1. OffType=Rape 44 ==> Gender=M 44    conf:(1)
 2. OffType=Sex_Assault 41 ==> Gender=M 41    conf:(1)
 3. OffType=Rape CriminalBkgd=Yes 40 ==> Gender=M 40    conf:(1)
 4. OffType=Sex_Assault CriminalBkgd=Yes 38 ==> Gender=M 38    conf:(1)
 5. OffDay=Sat CriminalBkgd=Yes 56 ==> Gender=M 55    conf:(0.98)
 6. OffDay=Sat Race=Black CriminalBkgd=Yes 34 ==> Gender=M 33    conf:(0.97)
 7. Race=Black CriminalBkgd=Yes 125 ==> Gender=M 120    conf:(0.96)
 8. OffDay=Sun CriminalBkgd=Yes 46 ==> Gender=M 44    conf:(0.96)
 9. OffDay=Sat 67 ==> Gender=M 64    conf:(0.96)
10. OffDay=Sat Race=Black 41 ==> Gender=M 39    conf:(0.95)
```

Apriori Output with 5 attributes.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3,9-weka.filters.unsupervised.attribute.Remove-R6
Instances:    330
Attributes:   5
              OffDay
              OffCounty
              OffType
              Gender
              Race
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (33 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 2

Best rules found:

 1. OffType=Rape 44 ==> Gender=M 44    conf:(1)
 2. OffType=Sex_Assault 41 ==> Gender=M 41    conf:(1)
 3. OffDay=Sat 67 ==> Gender=M 64    conf:(0.96)
 4. OffDay=Sat Race=Black 41 ==> Gender=M 39    conf:(0.95)
 5. OffDay=Sun 56 ==> Gender=M 53    conf:(0.95)
 6. Race=Black 162 ==> Gender=M 153    conf:(0.94)
 7. Race=Hspnc 51 ==> Gender=M 48    conf:(0.94)
 8. OffCounty=Monongalia Race=Black 42 ==> Gender=M 39    conf:(0.93)
 9. OffCounty=Monongalia 80 ==> Gender=M 74    conf:(0.93)
10. OffDay=Mon 39 ==> Gender=M 36    conf:(0.92)
```

Tertius Output with 6 attributes.

```
=== Run information ===

Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3,9
Instances:    330
Attributes:   6
              OffDay
              OffCounty
              OffType
              Gender
              Race
              CriminalBkgd
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.393963 0.030303 */ OffType = Accident ==> OffDay = Wed or OffCounty = Harrison or CriminalBkgd = No
 2. /* 0.392162 0.042424 */ OffType = Accident ==> OffCounty = Harrison or CriminalBkgd = No
 3. /* 0.385031 0.036364 */ OffType = Accident ==> OffDay = Mon or OffCounty = Harrison or CriminalBkgd = No
 4. /* 0.385031 0.036364 */ OffType = Accident ==> OffCounty = Harrison or Gender = F or CriminalBkgd = No
 5. /* 0.373996 0.039394 */ OffType = Accident ==> OffCounty = Harrison or Race = AorPI or CriminalBkgd = No
 6. /* 0.370933 0.109091 */ CriminalBkgd = No ==> OffType = Accident or Race = AorPI
 7. /* 0.367353 0.030303 */ OffType = Accident ==> OffDay = Tue or OffCounty = Harrison or CriminalBkgd = No
 8. /* 0.365576 0.045455 */ OffType = Accident ==> OffDay = Wed or CriminalBkgd = No
 9. /* 0.362788 0.096970 */ CriminalBkgd = No ==> OffCounty = Harrison or OffType = Accident or Race = AorPI
10. /* 0.362029 0.051515 */ OffType = Accident ==> OffDay = Mon or CriminalBkgd = No
11. /* 0.359633 0.060606 */ OffType = Accident ==> CriminalBkgd = No

Number of hypotheses considered: 53123
Number of hypotheses explored: 25257
Time: 00 min 19 s 859 ms
```

Tertius Output with 6 attributes with filter.

```
Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-weka.filters
Instances:    330
Attributes:   6
              OffDay
              OffCounty
              OffType
              Gender
              Race
              Age
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.347445 0.042424 */ OffType = Drugs ==> OffDay = Thru or Race = AorPI or Age = '(-inf-19.3]'
 2. /* 0.326950 0.030303 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Race = AorPI
 3. /* 0.320195 0.033333 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Age = '(-inf-19.3]'
 4. /* 0.319129 0.057576 */ OffType = Drugs ==> OffDay = Thru or Age = '(-inf-19.3]'
 5. /* 0.317121 0.048485 */ OffType = Drugs ==> OffDay = Thru or Gender = F or Age = '(-inf-19.3]'
 6. /* 0.317032 0.057576 */ OffCounty = Monongalia ==> OffDay = Sat or OffType = Sex_Assault or Age = '(19.3-24.6]'
 7. /* 0.312442 0.039394 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia
 8. /* 0.310521 0.027273 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Age = '(24.6-29.9]'
 9. /* 0.309606 0.033333 */ OffType = Burglary ==> OffDay = Wed or OffCounty = Cabell or Race = Hspnc
10. /* 0.303918 0.033333 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Gender = F

Number of hypotheses considered: 175531
Number of hypotheses explored: 87890
Time: 01 min 01 s 515 ms
```

Tertius Output with 5 attributes.

```
=== Run information ===

Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1-3,9-weka.filters.unsupervised.attribute.Remove-R6
Instances:    330
Attributes:   5
              OffDay
              OffCounty
              OffType
              Gender
              Race
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.326950 0.030303 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Race = AorPI
 2. /* 0.312442 0.039394 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia
 3. /* 0.309606 0.033333 */ OffType = Burglary ==> OffDay = Wed or OffCounty = Cabell or Race = Hspnc
 4. /* 0.303918 0.033333 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Gender = F
 5. /* 0.296739 0.084848 */ OffCounty = Monongalia ==> OffDay = Sat or OffType = Sex_Assault or Race = AorPI
 6. /* 0.295875 0.054545 */ OffType = Drugs ==> OffDay = Thru or Gender = F or Race = AorPI
 7. /* 0.294842 0.021212 */ OffDay = Thru ==> OffCounty = Kanawha or OffType = Drugs
 8. /* 0.289435 0.030303 */ OffType = Burglary ==> OffDay = Tue or OffCounty = Cabell or Race = Hspnc
 9. /* 0.284424 0.054545 */ OffType = Accident ==> OffDay = Mon or OffCounty = Harrison or Race = White
10. /* 0.281752 0.030303 */ OffDay = Thru ==> OffType = Drugs

Number of hypotheses considered: 45140
Number of hypotheses explored: 23205
Time: 00 min 15 s 985 ms
```

Tertius Output with 4 atributes.

```
=== Run information ===

Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     TestData-weka.filters.unsupervised.attribute.Remove-R1,10-weka.filters.unsupervised.attribute.Discretize-B10-
Instances:    330
Attributes:   4
              OffDay
              OffCounty
              OffType
              Race
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0.326950 0.030303 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia or Race = AorPI
 2. /* 0.312442 0.039394 */ OffType = Drugs ==> OffDay = Thru or OffCounty = Monongalia
 3. /* 0.309606 0.033333 */ OffType = Burglary ==> OffDay = Wed or OffCounty = Cabell or Race = Hspnc
 4. /* 0.296739 0.084848 */ OffCounty = Monongalia ==> OffDay = Sat or OffType = Sex_Assault or Race = AorPI
 5. /* 0.294842 0.021212 */ OffDay = Thru ==> OffCounty = Kanawha or OffType = Drugs
 6. /* 0.289435 0.030303 */ OffType = Burglary ==> OffDay = Tue or OffCounty = Cabell or Race = Hspnc
 7. /* 0.284424 0.054545 */ OffType = Accident ==> OffDay = Mon or OffCounty = Harrison or Race = White
 8. /* 0.281752 0.030303 */ OffDay = Thru ==> OffType = Drugs
 9. /* 0.279918 0.090909 */ OffCounty = Monongalia ==> OffDay = Sat or OffType = Rape or Race = AorPI
10. /* 0.277936 0.066667 */ OffType = Drugs ==> OffDay = Thru or Race = AorPI

Number of hypotheses considered: 23697
Number of hypotheses explored: 12120
Time: 00 min 07 s 140 ms
```