Portfolio Milestone

Jonah Witt

Syracuse University

September 17, 2019

<https://github.com/jwwitt/Jonah_Witt_Portfolio_Milestone>

Table of Contents

[1. Introduction 3](#_Toc19458678)

[2. IST 659: Database Administration 4](#_Toc19458679)

[a. Project Description 4](#_Toc19458683)

[b. Learning Goals 6](#_Toc19458684)

[3. IST 707 Data Mining 6](#_Toc19458682)

[a. Project Description 7](#_Toc19458683)

[b. Learning Goals 8](#_Toc19458684)

[4. IST 736 Text Mining 8](#_Toc19458685)

[a. Project Description 9](#_Toc19458686)

[b. Learning Goals 11](#_Toc19458687)

[5. IST 718: Big Data Analytics 11](#_Toc19458688)

[a. Project Description 12](#_Toc19458689)

[b. Learning Goals 12](#_Toc19458690)

[6. Conclusion 14](#_Toc19458691)

[7. References 16](#_Toc19458692)

**1. Introduction**

The MS in Applied Data Science from the School of Information Studies at Syracuse University is designed to emphasize the application of data science to enterprise operations and processes, particularly in the areas of data capture, management, analysis, and communication for decision-making. The program accomplishes this by implementing the following learning goals:

1. Describe a broad overview of the major practice areas in data science
2. Collect and organize data
3. Identify patterns in data via visualization, statistical analysis, and data mining
4. Develop alternative strategies based on the data
5. Develop a plan of action to implement the business decisions derived from the analyses
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy)

These learning objectives goals were achieved in projects completed throughout the ADS program. In particular, projects completed in IST 659: Database Administration (Witt, “IST 659”), IST 707: Data Mining (Witt, “IST 707”), IST 736: Text Mining (Witt, “IST 736”), and IST 718: Big Data Analytics (Witt, “IST 718”) successfully demonstrated the seven learning goals for the ADS program.

**2. IST 659: Database Administration**

## **a. Project Description**

In IST 659: Database Administration, a database was created to store college football data from the 2017 football season (Witt “IST 659”). This database contained data on players, coaches, teams, conferences, and games. These data and statistics were gathered from the internet. The database was initially developed by creating a conceptual and logical model of the Database Diagram (Fig. 1). This ensured that all of the data was in the correct format and the relationships between entities were properly defined. Next, the database was created in Microsoft SQL Server. The database was created to only allow authorized users to be able to edit and access the data in the data tables. In addition to the data tables, several queries and procedures were created to allow information to be easily extracted from the database.

Once the database was up and running on the server, a User Interface was developed using Microsoft Access (Fig. 2). The user interface allowed for authorized users to easily add and query the data. A menu was created for data entry, and additional menus were created for querying data on players, coaches, teams, games, and seasons. When the user makes a selection of a query, the user is prompted to enter the values that the data will be pulled from. The output of the query is then presented to the user (Fig. 3).

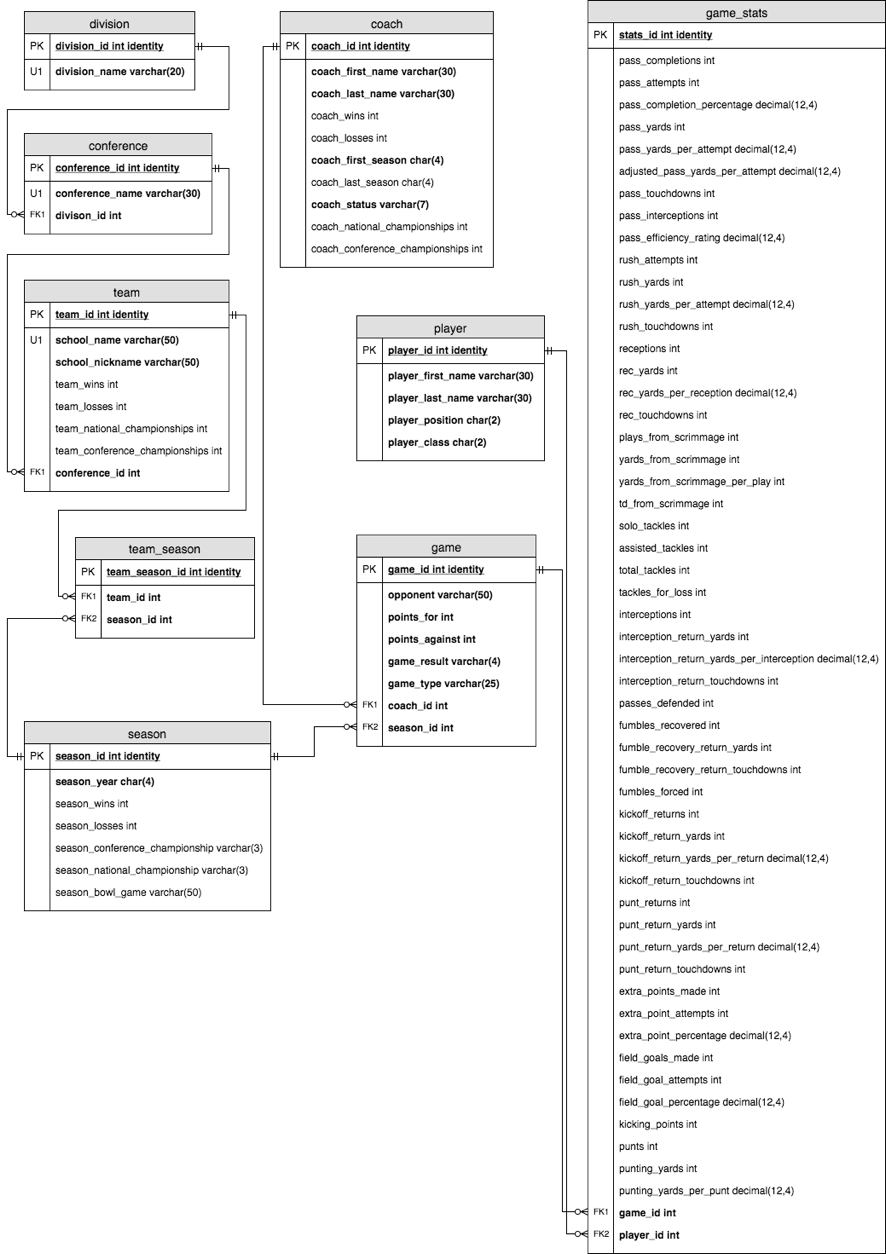


Figure 1: Database Diagram (Logical Model)

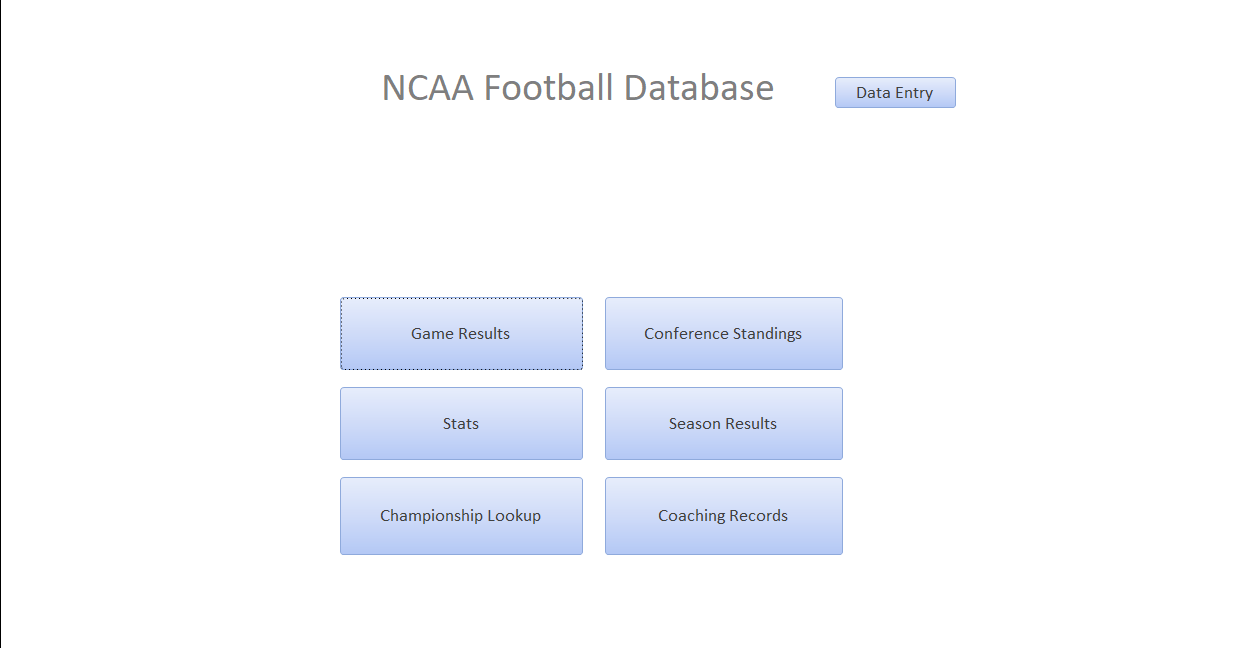


Figure 2: Database User Interface

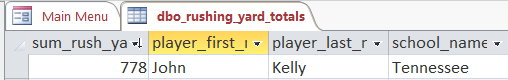


Figure 3: Example Query Results

## **b. Learning Goals**

The development of the College Football Database demonstrated the ability to collect and organize data, develop alternative strategies based on the data, and synthesize the ethical dimensions of data science practice. The gathering of data from the website along with the structure developed in the database diagram display the ability to collect and organize data. Alternative strategies were developed during the creation of the database. Initially, the database was going to contain data across multiple college football seasons. However, the amount of data was too large, so the scope of the project was limited to the 2017 football season. Lastly, the ethical dimensions of data science had to be considered when creating the database. While the data used in the project is public, important security measures had to be implemented in order to ensure that the data was only accessed by those that were authorized. These same practices could be implemented to more sensitive databases (credit cards, social security numbers, etc.) in a similar fashion. The concepts demonstrated in this project lay the foundation for database collection and organization in the professional setting. These concepts will be built upon in the final semester through IST 722: Data Warehousing.

**3. IST 707 – Data Mining**

## **a. Project Description**

In IST 722: Data Mining, data from the FIFA world cup was analyzed to determine the economic impact of the world cup (Witt, “IST 722”). All of the data analysis was conducted using R in R Studio. The data was analyzed from a data set from Kaggle.com. More specifically, two data sets from Kaggle were analyzed. One data set contained in-game match data, and another data set contained data on each independent World Cup. Using R, the data was processed, visualized, and several machine learning algorithms were used to predict the outcome of a world cup based on the features provided. The machine learning algorithms implemented were Association Rules (Fig. 4), K-Means Clustering, K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Random Forest. These models, combined with economic data gathered separately, determined that it is not economically beneficial to host the world cup. The biggest economic gain to be made from the world cup came from prize money.

**Association Rules**

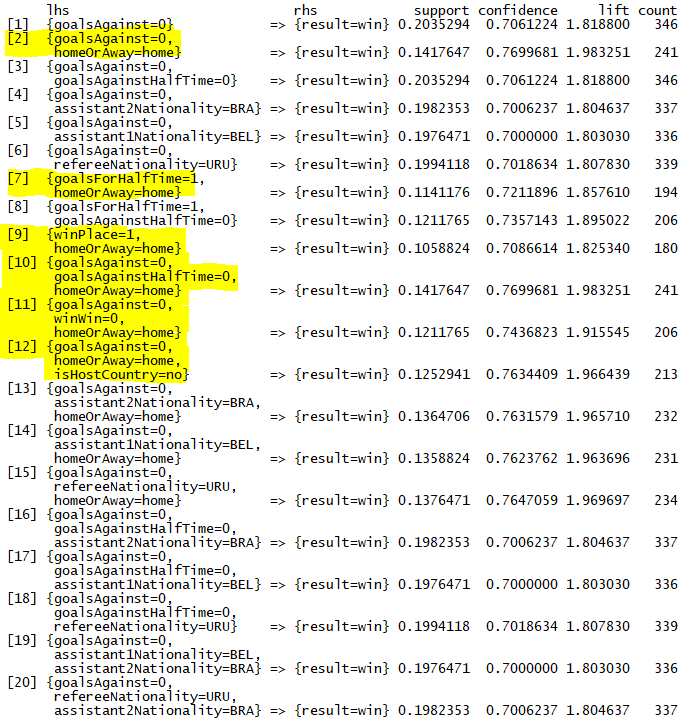


Figure 4: Association Rule Analysis

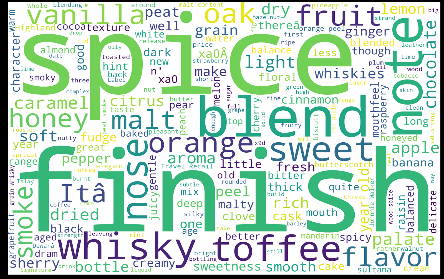
## **b. Learning Goals**

The analysis of the FIFA World Cup data demonstrated the ability to identify patterns in data via statistical analysis and data mining, develop alternative strategies based on the data, and demonstrate communication skills regarding data and its analysis. Patterns were identified in the data using data mining. In particular, association rule analysis identified several patterns that lead to a team winning a game. Additionally, alternative strategies had to be developed based on the data. The data provided showed early on that hosting the world cup was not economically beneficial unless the world cup was won. Because of this fact, the data mining tools were leveraged to develop models to predict the likelihood that a team wins the world cup. Lastly, communication skills regarding data and its analysis were demonstrated in both the written project report and the final presentation. The presentation was conducted as if it were being given to a nation who is deciding whether or not to host the world cup. The data and results were conveyed in the presentation to allow the nation to make the best decision. The tools used in this project can be applied in the professional setting to identify patterns in data and develop models to predict results.

**4. IST 736: Text Mining**

## **a. Project Description**

In IST 736: Text Mining, reviews on different brands of scotch were analyzed in order to predict the sentiment of a review from its content (Witt, “IST 736”). The data analysis was done in Python, particularly the Scikit-Learn package. The scotch reviews came from a Kaggle data set, but they were initially from “Whiskey Advocate”. In order for the review text to be converted into features for machine learning algorithms, a feature was created for each word based on its frequency in the review. These features were visualized using word clouds (Fig. 5).



*Overall Word Cloud Blended Malt Scotch Whisky Blended Scotch Whisky*

Figure 5: Word Clouds for Whisky Reviews

Once the data was cleaned, several machine learning algorithms were implemented in order to classify the sentiment of each review. The algorithms used were k-Means Clustering (Fig. 6), Naïve Bayes, LDA Topic Modeling, Bernoulli, and Support Vector Machine.

**k-Means Clustering**

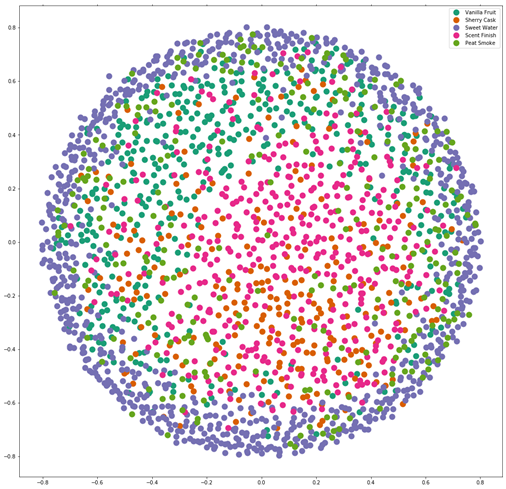


Figure 6: K-means Clustering for Whisky Reviews

The algorithms, particularly the support vector machine, determined which words were most influential in determining positive or negative sentiment in a review (Fig. 7). These results were conveyed in the presentation and could be used as recommendations to the marketing team of a scotch manufacturer.

**Most Influential Words for Sentiment**

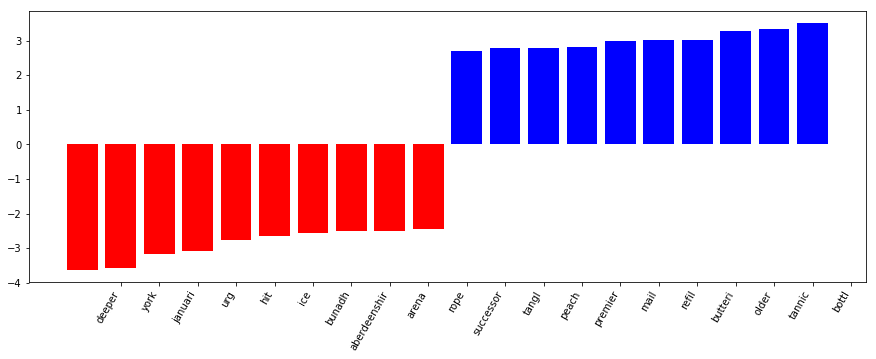


Figure 7: Feature Importance for Whisky Reviews

## **b. Learning Goals**

The analysis of the scotch reviews demonstrated the ability to identify patterns via visualization, statistical analysis and data mining, develop a plan of action to implement the business decisions derived from the analyses, demonstrate communication skills regarding data and its analyses and synthesize the ethical dimensions of data science practice. The ability to identify patterns was demonstrated through both the visualization of the reviews and the results of the machine learning algorithms. The models were able to determine which features were most influential on the sentiment of a review. These patterns were then used to communicate a plan to the marketing team of scotch manufacturers in order to gain better online review ratings. The communication skills were demonstrated in both the written report and the final presentation of results. Lastly, the ethical dimensions of data science had to be analyzed to determine which features were included from the reviews. All user information was excluded in order to remove any potential bias in the models. This project provided the building blocks for gaining information from unstructured data. These tools can be used in the professional setting to extract data from sources that would otherwise be unable to be analyzed.

**5. IST 718: Big Data Analytics**

## **a. Project Description**

In IST 718: Big Data Analytics, twitter data was analyzed in order to classify tweets as either “Hateful” or “Not Hateful”. The data set, containing 30,000 tweets, was from Analytics Vidhya. Text mining concepts were used to create a data processing pipeline for the tweets (Fig. 8).

**Data Processing Pipeline**

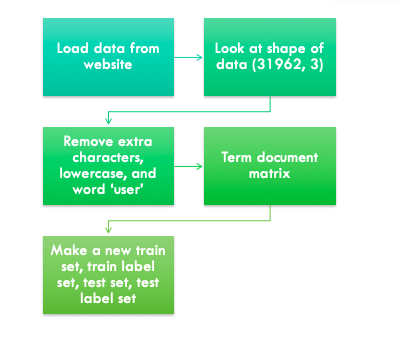


Figure 8: Data Cleaning Process for Twitter Data

Once the tweets were processed, a classifier was tested using several machine learning algorithms. The algorithms tested for the classifier were Naïve Bayes, Random Forest, Support Vector Machine, and Neural Networks (Keras). While the neural network created the best classifier, the support vector machine was much more computationally efficient. It was recommended that the support vector machine be used to continuously train the classifier to identify hate speech. The feature importance (Fig. 9) and confusion matrix (Fig. 10) from the support vector machine are shown below.

**Influential Words for Determining Hate Speech**



Figure 9: Feature Importance for Tweet Sentiment

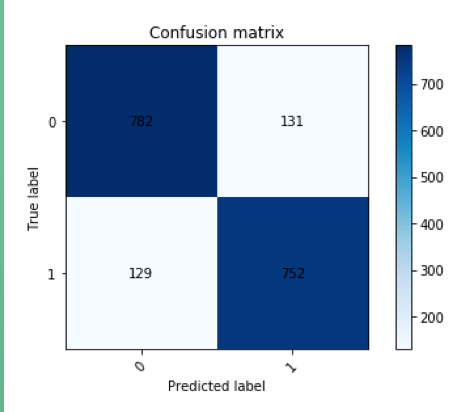


Figure 10: Confusion Matrix for Support Vector Machine

**b. Learning Goals**

The development of a classifier to classify tweets as “Hateful” or “Not Hateful” demonstrated the ability to identify patterns in data via statistical analysis and data mining, develop alternative strategies based on the data, develop a plan of action to implement the business decisions derived from the analyses, and communicate the data and analyses. The ability to identify patterns in the data was demonstrated by finding the influential features to determine if a tweet is “Hateful” or “Not Hateful”. Alternative strategies were made based on the data. The initial plan was to use the keras neural network model to implement the classifier. However, the most influential features showed that the trainer would need to be continuously trained to account for current events. Because of this fact, the support vector machine was chosen for its computational efficiency. Additionally, a plan of action was developed to create the classifier and routinely train the classifier with new data. This will allow the classifier to identify hateful speech derived from current events and new slang. Lastly, communication skills were once again demonstrated in both the written report and final presentation. The skills learned in this project can be applied in the professional setting by developing data pipelines and complex machine learning models.

**6. Conclusion**

The Applied Data Science program at Syracuse University prepares students to solve data problems in the real world by implementing seven learning goals. The projects in this portfolio successfully demonstrate all seven of the learning goals of the ADS program. Additionally, the summation of all four projects describes a broad overview of the major practice areas in data science. The ability to collect and organize data is demonstrated in IST 659: Database Administration (Witt, “IST 659”). The ability to identify patterns in data via visualization, statistical analysis, and data mining is demonstrated in IST 707: Data Mining (Witt, “IST 707”), IST 736: Text Mining (Witt, “IST 736”), and IST 718: Big Data Analytics (Witt, “IST 718”). The ability to develop alternative strategies based on the data was demonstrated in IST 707: Data Mining and IST 718: Big Data Analytics. A plan of action to implement the business decisions derived from the analyses was developed in IST 736: Text Mining and IST 718: Big Data Analytics. Communication skills regarding data and its analyses for managers, IT professionals, programmers, statisticians, and other relevant professionals were demonstrated in IST 707: Data Mining, IST 736: Text Mining, and IST 718: Big Data Analytics. Lastly, the ability to synthesize the ethical dimensions in data science practice was demonstrated in IST 736: Text Mining and IST 718: Big Data Analytics. These four projects have successfully demonstrated the skills necessary to have success implementing data science practices in the professional domain.

# **7. References**

Witt, Jonah. (2018). IST 659: Final Project. Unpublished manuscript, Syracuse University.

Witt, Jonah. (2019). IST 707: Final Project. Unpublished manuscript, Syracuse University.

Witt, Jonah. (2019). IST 736: Final Project. Unpublished manuscript, Syracuse University.

Witt, Jonah. (2019). IST 718: Final Project. Unpublished manuscript, Syracuse University.