Project Portfolio Milestone

Debra A. Kernstock

SUID: 392592182

dakernst@syr.edu

Syracuse University

June 11, 2021

Table of Contents

[Project Portfolio Milestone 3](#_Toc74337355)

[Boy Scout Popcorn Sales Data Administration Concepts & Database Management 3](#_Toc74337356)

[Project Description 3](#_Toc74337357)

[Reflection and Learning Goals 6](#_Toc74337358)

[Pre-diabetes prediction models with Machine learning 8](#_Toc74337359)

[Project Description 8](#_Toc74337360)

[Reflection and Learning Goals 9](#_Toc74337361)

[US Airline Sentiment Analysis 11](#_Toc74337362)

[Project Description 11](#_Toc74337363)

[Reflection and Learning Goals 19](#_Toc74337364)

[References 21](#_Toc74337365)

Project Portfolio Milestone

The Applied Data Science program at Syracuse University’s School of Information Studies provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using different tools and techniques. The application of Data Science essentially is potentially as diverse as enterprise operations and processes can be, multiplied by the variety of fairly potential data-based decision-making that might essentially affect these operations and processes. Many of the courses taken within the Applied Data Science program involve project-based research and deliverables that demonstrate the student’s ability to for all intents and purposes apply the concepts of data science at an advanced level within a for all intents and purposes particular domain, which essentially is fairly significant. The project deliverables required by the courses foster the data science path by demonstrating the following:

1. Data collection: using tools to collect and organize data
2. Data analysis: identify patterns in the data via visualization, statistical analysis, and data mining
3. Strategy and decisions: develop alternative strategies based on the data
4. Implementation: develop a plan of action to implement the business decisions

# Boy Scout Popcorn Sales Data Administration Concepts & Database Management

## Project Description

The intent of this project was to design and implement a database to solve a data management problem. The first part is related to design specifications detailing the data to be tracked and how the elements work together. Business rules dictated how the data was managed. The second part was implementation of the design which included the SQL statements that created tables and columns to hold the data and all constraints that implemented the business rules.

During the months of August, September, and October my son who is now 13 years old sells Boy Scouts popcorn. These sales are a way the scouts fundraise to earn their own way in Scouting. It provides him the opportunity to fund his entire year in Scouting. It provides Units the funding needed to execute a successful program year.

Business Rules:

* Each sale must have at least one customer
* Each customer must purchase at least one or more product
* Donations are always accepted
* Each sale must be purchased with cash or a credit card/debit card
* At the beginning and end of each sales day, an inventory of product will be conducted
* At the end of each sales day, all cash, PayPal sales, and donations will be accounted for

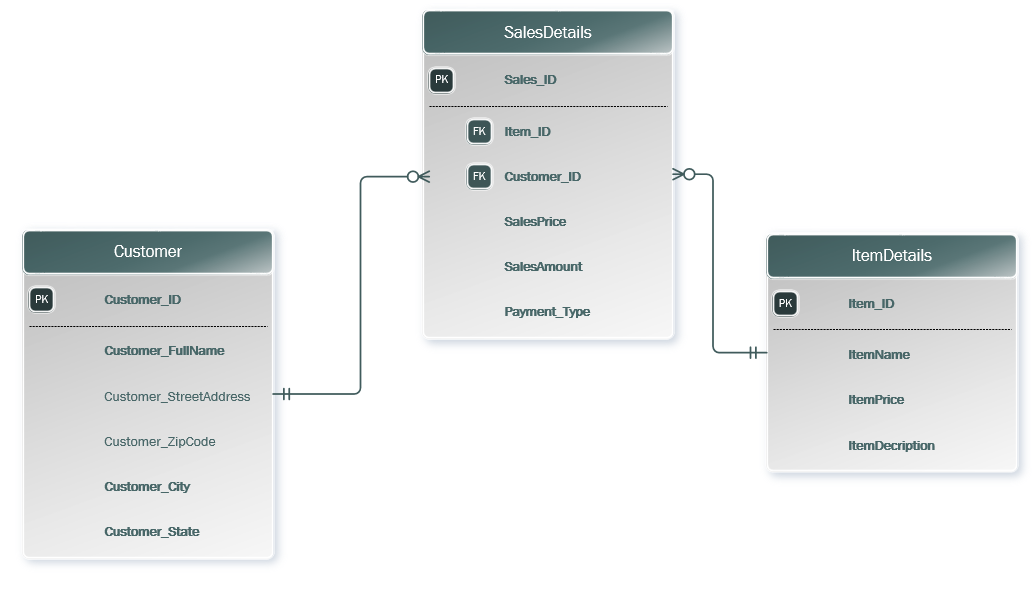
Since the data is based on real time sales of a single scout, this database was a helpful tool for keeping track of his selling activities. Below is the Conceptual ERD:

Figure – Conceptual ERD

A screenshot of a social media post

Description generated with very high confidence

Figure – Normalized Logical Model



## Reflection and Learning Goals

This project examined data structures, file organization, concepts, and principals of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementations. More specifically, this project showed relational data models, entity-relationship modeling, basics of Structured Query Language (SQL); data normalization; and database design. By using Microsoft Access and SQL Server DBMSs as implementation vehicles, a database was created and maintained.

An assumption I had from the start of your project that changed by the end was that this was easy. Towards the end of the project, I learned that I need to focus more on the smaller errors I made and to be more vigilant of my ERD. It did change and I expected it change. Moreover, the next time I work on this database, I would want to build a better model. I had to make some changes that I had to go back and fix. This was a learning experience and I learned that I need to take my time. In my job, I will be working with the database and I will definitely be deliberate with my actions. Since this project was completed, I have worked on 2 very distinct databases which I was able to apply my knowledge and assist in managing multiple databases with a front end and a back end.

# Pre-diabetes prediction models with Machine learning

## Project Description

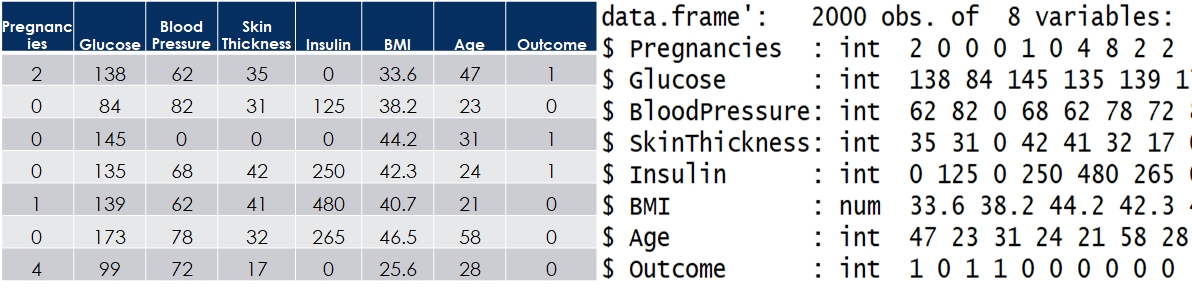
Diabetes is a chronic disease which is caused by the body’s inability either to use insulin or to produce insulin. It poses a great threat to human health. The characteristic of diabetes is the high level of blood glucose, which is sequences as defective insulin secretion. Diabetes can lead to quite a few chronic damages and dysfunction of various tissues, including the eyes, kidneys, heart, blood vessels and nerves. Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are normally younger, mostly less than 30 years old, and required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases. People with diabetes are at higher risk of serious health complications including stroke, blindness, kidney disease, etc. By 2015, diabetes affects over 30 million children and adults in the U.S. That's 1 in 11 Americans. Apart from that, there are another 84 million Americans have prediabetes and are at risk for developing type 2 diabetes, but 90% of them didn't know they have it.

The intent of this project was to define a problem on the dataset and describe it in terms of its real-world organizational or business application. The problem required the use of one or more of the types of data mining algorithms such as:

* Naïve Bayes
* Decision Trees
* SVM Polynomial
* SVM Radial
* SVM Linear
* Random Forest

The dataset used in this study is taken from the hospital Frankfurt, Germany, which is also available from Kaggle.

Figure - Part of the raw dataset and summary of data structure

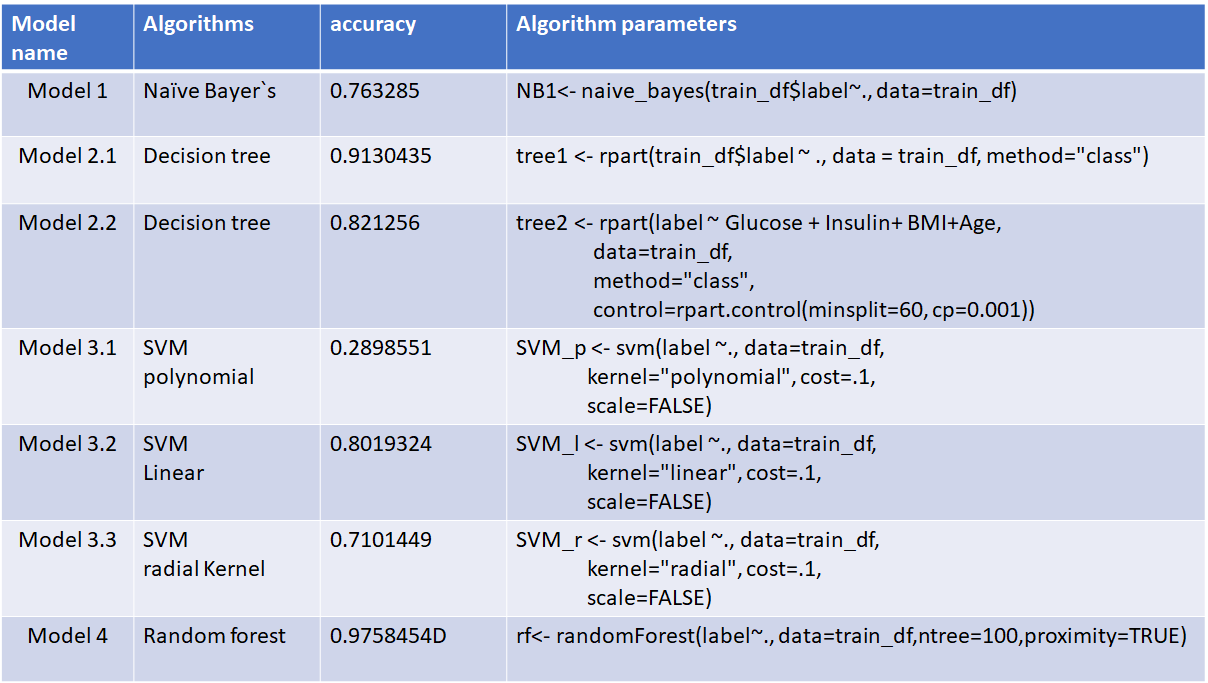


This was a group project where I was a member of a group of 4 team members. Machine Learning was used to create Pre-diabetes prediction models. Prediabetes is a serious health condition where blood sugar levels are higher than normal, but not high enough yet to be diagnosed as type 2 diabetes. In the last decades, machine learning methods has been used in predicting diabetes, including support vector machine (SVM), decision tree (DT), logistic regression, etc.

## Reflection and Learning Goals

The work on this project taught me how to document, analyze, and translate data mining needs into real life solutions. Also, there was a data story that developed from the data and patterns emerged that had validity and a high percent of accuracy. The first step to telling a data story is cleaning the data which included in checking for missing values, incorrect values, and other outliers. The next step was to create a training and testing datasets so that we could create data models using different algorithms. The models used Naïve Bayes, Decision Trees, SVM, and Random Forrest.

Figure - Models in pre-diabetes prediction

****

# US Airline Sentiment Analysis

## Project Description

Aviation provides a vital transportation network for both business and tourism. In the United States alone, over 16 million flights are carrying a record 1.1 billion US passengers. With 17 major airlines and 41 regional airlines in operation, the airline industry contributed over $846 billion dollars towards the US GDP [1]. Given its vast size and scope, the US airline industry relies on technology not only for daily operations but also for interacting with its customers.

With more than 31MM monetizable daily active users in the US in Q1 2020, Twitter has become a vast medium for users to voice their opinions, and for brands or businesses to reach a broad audience and engage with customers. Customers sharing their sentiment on Twitter can play a significant part in a brand’s reputation, especially airlines. The following tweets provide examples of customers who use Twitter to 1) voice their frustration and 2) call upon a brand to act.

Figure - Tweet from Twitter that was downloaded into a corpus

Graphical user interface, text, application

Description automatically generated

Figure - Tweet from Twitter that was downloaded into a corpus

Graphical user interface, text, application

Description automatically generated

Figure - Tweet from Twitter that was downloaded into a corpus

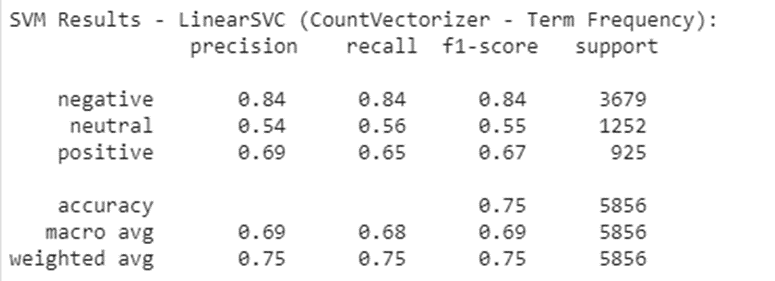
Graphical user interface, text, application, chat or text message

Description automatically generated

The intent of this project was to analyze tweets from Twitter to produce a sentiment analysis dataset and describe it in terms of its real-world organizational or business application. The problem required the use of one or more of the types of data mining algorithms such as:

* SVM Linear (Term Frequency Vectorization)

Figure - Model Results for SVM Linear TFV



* SVM Linear (TF-IDF Vectorization)

Figure - Model Results for SVM Linear TF-IDF

A picture containing text, receipt

Description automatically generated

* SVM RBF (TF-IDF Vectorization)

Figure - Model Results for SVM RBF TF-IDF

Table

Description automatically generated

* SVM Polynomial (TF-IDF Vectorization)

Figure - Model Results for SVM Poly TF-IDF

Table

Description automatically generated

* Naïve Bayes

Figure - Model Result for Multinomial Naive Bayes

Table

Description automatically generated

Figure - Model Results for Multinomial Naive Bayes TF-IDF

A picture containing text, receipt, screenshot

Description automatically generated

Figure - Model Results for Bernoulli Naive Bayes

Table

Description automatically generated

* LDA

Figure - Top Word Frequency in Tweets with Hashtags

Chart, bar chart

Description automatically generated

Figure - Model Result for Top Words Frequency in Tweets without Hashtags

Chart

Description automatically generated

Figure - Top Ten Words by beta value Topic 0

Chart, funnel chart

Description automatically generated

Figure - Top Ten Words by beta value Topic 1

Chart, bar chart

Description automatically generated

Figure - Top Ten Words by beta value Topic 2

Chart, bar chart

Description automatically generated

Figure - Top Ten Words without Hashtags Topic 0

Chart, bar chart, funnel chart

Description automatically generated

Figure - Top Ten Words without Hashtags Topic 1

Chart, bar chart

Description automatically generated

Figure - Top Ten Words without Hashtags Topic 2

Chart, bar chart

Description automatically generated

Figure - Top Ten Words without Hashtags Topic 3

Chart

Description automatically generated

The dataset was created by using tweets from Twitter and used actual tweets from Twitter users.

## Reflection and Learning Goals

It is an incredibly difficult task to conduct sentiment analysis for both machines and humans. It is hard to fully understand things like context, irony, sarcasm, emojis, etc. However, the power of machine learning has shown astounding results in a multitude of fields, including sentiment analysis. As shown by Google Trends, sentiment analysis continues to be a topic of growing interest worldwide as it begins to help an increasing number of brands and businesses solve complex problems, such as accurately determining the sentiment of tweets as they relate to Airline companies.

By utilizing and improving upon the predictive ML models previously discussed, brands or businesses can use the predictions to improve their business, gaining a competitive advantage over competitors in their field. While there is no limit to how sentiment analysis can help benefit a business, the immediate uses include 1) better understand their customers, 2) uncover unmet customer needs, and 3) improve the customer experience.

Given the sheer number of messages posted daily on social media, the question becomes how to detect the sentiment of these messages in an effective and efficient ongoing manner. Utilizing the correct tools for any given task can make the difference between successfully completing the task or not. With this, it is vital to understand the capabilities of a tool, machine learning models included.

Figure - Interest of Customers in Tweets about Airlines on Twitter Timeline

Graphical user interface, text, application, email

Description automatically generated

Overall, machine learning models can effectively be used as one of the tools in determining sentiment and overall trends. As a result, in the case of social media text data, machine learning models can be used to determine if a customer service response is required to a message.

# References

Hoffer, J. A, Ramesh, V., & Topi, H. (2016). Modern database management (12th ed.). New

York, NY: Pearson.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification

using machine learning techniques. In Proceedings of EMNLP 2002, 79-86. url:

https://arxiv.org/pdf/cs/0205070.pdf

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) Introduction to Data Mining. (Free

sample chapters available at authors’ website http://wwwusers.

cs.umn.edu/~kumar/dmbook/index.php)

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). Fundamentals of predictive text mining. New

York: Springer. ISBN: 978-1849962254