Syracuse University, School of Information Studies

Master of Science in Applied Data Science

Portfolio Milestone

Kanning Wu

SUID: 714914874

[kwu130@syr.edu](mailto:kwu130@syr.edu)

The applied data science program at Syracuse University’s School of Information Studies provides students the opportunities to collect, analyze, decide and implement all data related tools and techniques. Each project listed below, data were collected from online sources and stored in SQL or .csv file. By using Python or R packages, visualization and analysis were generated and implemented. From the calculated results, insights from datasets can be developed. All these skills I learned at the School of Information Studies helped future data scientist focusing on the field of accounting, finance, marketing analytics.

Here is the list of projects:

Courses Project Overview Skills

IST 659 Database designing related to SQL Server, Access

COVID-19 Power BI,

SQL Server Data Tools

IST 687 Global Fossil Fuel Consumption R, Machine Learning

IST 707 Predicting Airbnb Prices in NYC R, Prediction Modeling,

Data Analysis

IST 736 Data Analysis of the City of San Python, Visualization,

Diego Data Analysis, Natural

Language Processing

Course Objectives:

IST 659

Describe fundamental data and database concepts

Explain and use the database development lifecycle

Create databases and database objects using popular database management system products

Solve problems by constructing database queries using Structured Query Language (SQL) Design databases using data modeling and data normalization techniques

Develop insights into future data management tool and technique trends

Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery Critique the effectiveness of DBMS in computer information systems

IST 687

Essential concepts and characteristics of data

Scripting/code development for data management using R and R-Studio

Principles and practices in data screening, cleaning, and linking

Communication of results to decision makers.

IST 707

Document, analyze, and translate data mining needs into technical designs and solutions.

Apply data mining concepts, algorithms, and evaluation methods to real-world problems.

Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

IST 736

Describe basic concepts and methods in text mining, for example document representation, information extraction, text classification and clustering, and topic modeling;

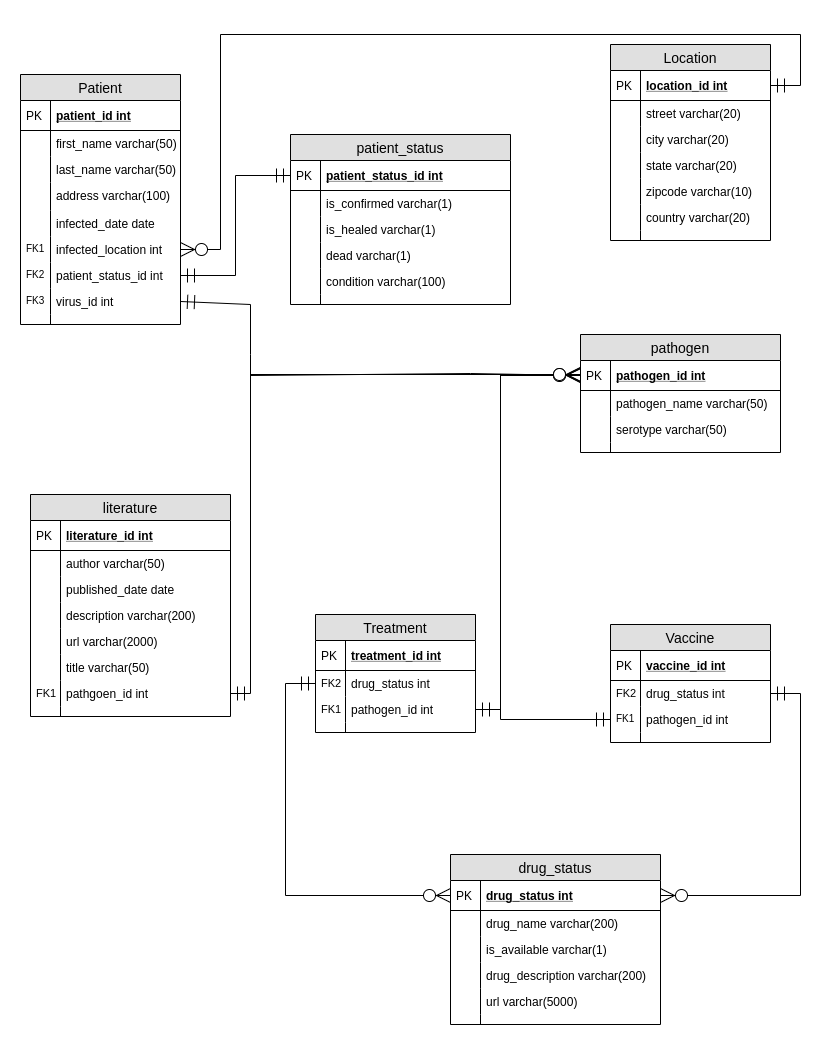
Use benchmark corpora, commercial and open-source text analysis and visualization tools to explore interesting patterns;

Understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems; and

Choose appropriate technologies for specific text analysis tasks and evaluate the benefit and challenges of the chosen technical solution.

IST 659 Final Project

This project is to design a database for COVID-19. It can be divided into following parts. 1. The virus information 2. Infected patients 3. Drugs availability 4. Locations 5. Journals. This database can provide information for people to keep tracking the current situation and understand the prevention of being infected. Also, this database is able to answer following questions: What is the current number of infected people? How many people are healed/dead? What location should be avoided going to? What are the methods can be used to prevent being infected? Are quarantined patients’ actions being tracked? Is the outbreak caused by virus, bacteria or other pathogens? DB schema is shown as:

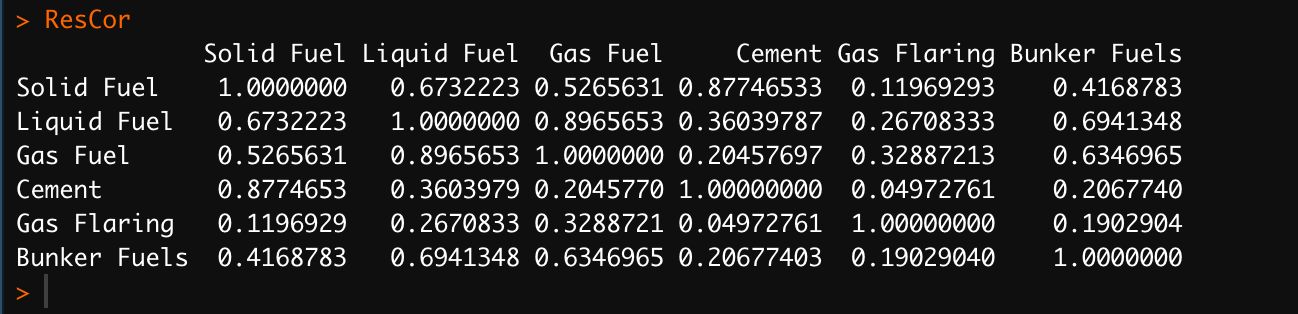


In this project, the database was implemented under development lifecycle. By identifying the issue and possible outcome, conceptual design was made, such as what tables and variables were needed. Since there are different types of pathogens, treatment, and prevention techniques, the design was not directly entered into MSSQL database management studio. Instead, it was first created on sketch paper and all the possible tables were normalized. While normalizing the database tables, a few designing issues were justified. When designing the database, there was some misunderstanding on how pandemic can occur and each pathogen can have multiple different treatments and preventions. The tables were redesigned. After implemented the database, there were some flaws on the connections between each table. The IDs of each table were mixed and errors were caused when adding sample data into the database. By correcting these errors, the database was finally can be used and helped to answer those questions required.

IST 687 Final Project

In this project, we applied the K-means clustering algorithm and Elbow Method, Silhouettes Methods to test the best number of clusters. For the K-means clustering algorithm, we first cleaned data and standardize it. Then, we tested the algorithm with different numbers of clusters from 2 to 10. For Elbow Method and Silhouettes Methods, we implemented both functions. With these two functions, we reached a result of having 10 as our optimal number of clusters. After we had the best number of clusters, we ran our code and got the final conclusion. The datasets we used were from datahub.com. Here are some visualizations:





This project applied essential concepts and characteristics of data to address real world problem. By using R and R-Studio, I had familiarized with this analytic language. From cleaning data and making visualizations, principle and practices in data screening and cleaning were the most important knowledge I learned throughout this course.

As IST 687 is the first class I took in this program. This project is not very complicated. Dataset was gathered from both Kaggle and a government website// The datasets contain categorical, dummies and numerical data. Since it is a project to answer imposing regulation of carbon emission in US and some details of the dataset were focusing on researching and other purposes. Some of the columns in the dataset were dropped by using R and R-Studio. Also, the dataset contains a lot of NAs which may cause inaccuracy of the results. These rows with NAs were also dropped or converted into 0s as needed. After implementing K-mean clustering, it is clear that US has more carbon emission than other developed countries, which can be provided to decision makers.

IST 707 Final Project

This project aims to create a model that can predict the price of an Airbnb using various classification methods. This model can be useful for customers to make sure they are paying a fair price or look for the best deals on the market. Using data from New York City’s Airbnb market and comparing different classification and prediction models, the most accurate housing price prediction of an Airbnb can be found. The models we are based on Naïve Bayes, R-part decision trees, Linear SVM, kNN and random forest models.

The dataset used to create these models from Kaggle consists of over 48,000 listed rentals spanning all five boroughs of New York city. The rentals each contain specific information ranging from the listers name to how many days the listing has been available.

From the dataset, there are a few visualizations created

**A screenshot of a cell phone

Description automatically generated**

**A sign lit up at night

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

After the visualization, classification models generated following results:

**R-Part Decision Tree**

The R-Part Decision Tree model predicted 78.57% of the test data correctly. It was more accurate at predicting listings that were below average, having a higher precision and recall for those listings. The final cp for the model was .024

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| R-Part | 78.57% | 86.9% | 83.1% | 59.7% | 66.5% |

**J48 Decision Tree**

The J48 Decision Tree model predicted 81.42% of the test data correctly. It was also more accurate at predicting listings that were below average, having a higher precision and recall for those listings. (seed = 1, numFolds = 5)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| J48 | 81.42% | 86.7% | 86.3% | 69% | 69.8% |

**Naïve Bayes:**

The Naïve Bayes model predicted 79.53% of the test data correctly. Similar to the previous two models, the NB model was more accurate in predicting listings that were below average. (with discretization, kernel estimation, and numFolds = 4, seed = 3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| Naïve Bayes | 79.53% | 83.4% | 88% | 68.6% | 59.9% |

**K-Nearest Neighbor**

The KNN correctly predicted 79.48% of the test data. It actually returned the same prediction measures as the Naïve Bayes model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| KNN | 79.48% | 83.4% | 88% | 68.6% | 59.9% |

**Random Forest**

The Random Forest model correctly predicted 80.96% of the test data correctly. Like every other model, the RF model is more accurate in predicting below-average prices. The final model was run with 10 trees.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| Random Forest | 80.96% | 86.7% | 85.8% | 68.2% | 69.9% |

**Model Summary and Comparisons (Green = Highest, Red = Lowest)**

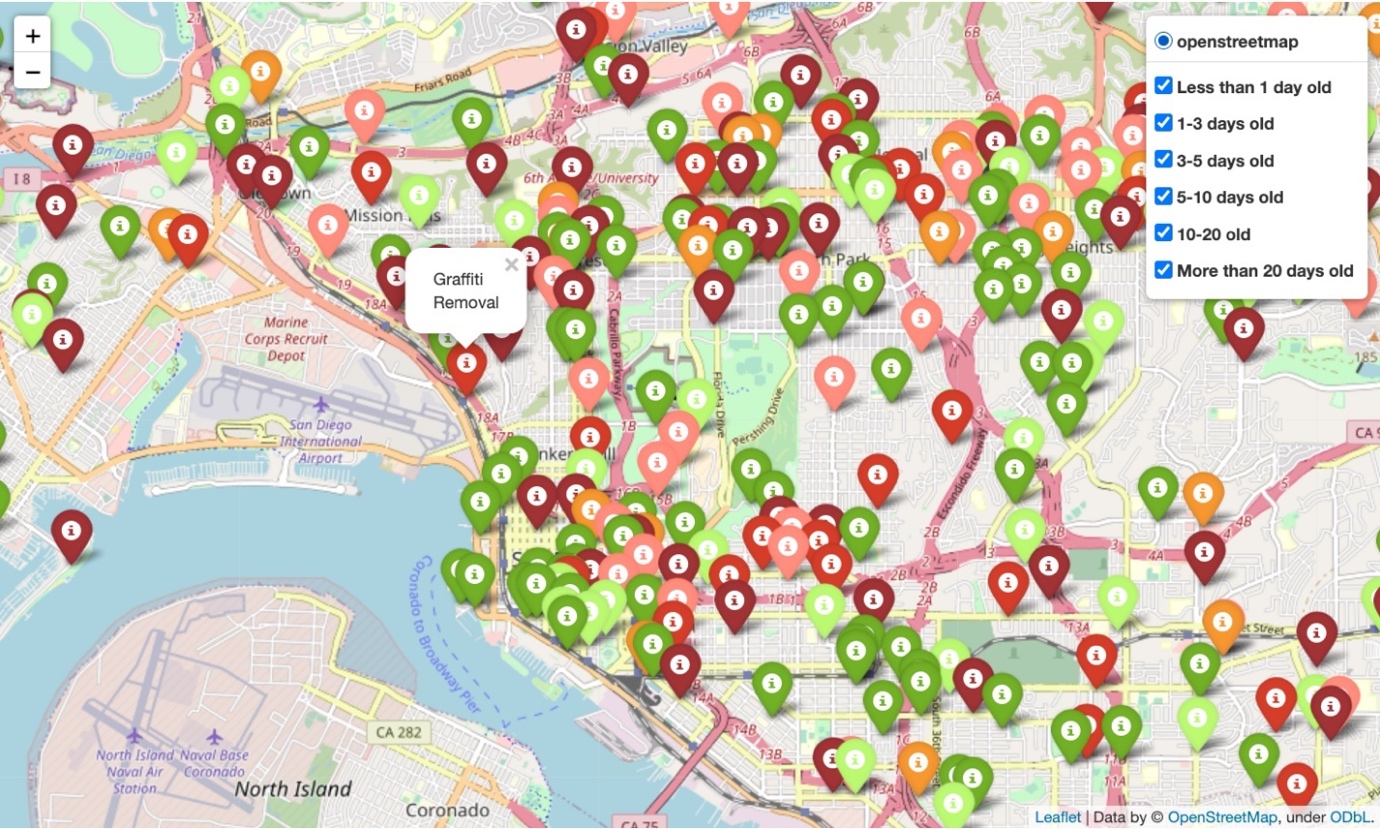
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Overall Accuracy | Precision <Avg | Recall <Avg | Precision >Avg | Recall >Avg |
| R-Part | 78.57% | 86.9% | 83.1% | 59.7% | 66.5% |
| J48 | 81.29% | 86.7% | 86.3% | 69% | 69.8% |
| Naïve Bayes | 79.48% | 83.4% | 88% | 68.6% | 59.9% |
| KNN | 79.48% | 83.4% | 88% | 68.6% | 59.9% |
| Random Forest | 80.96% | 86.7% | 85.8% | 68.2% | 69.9% |

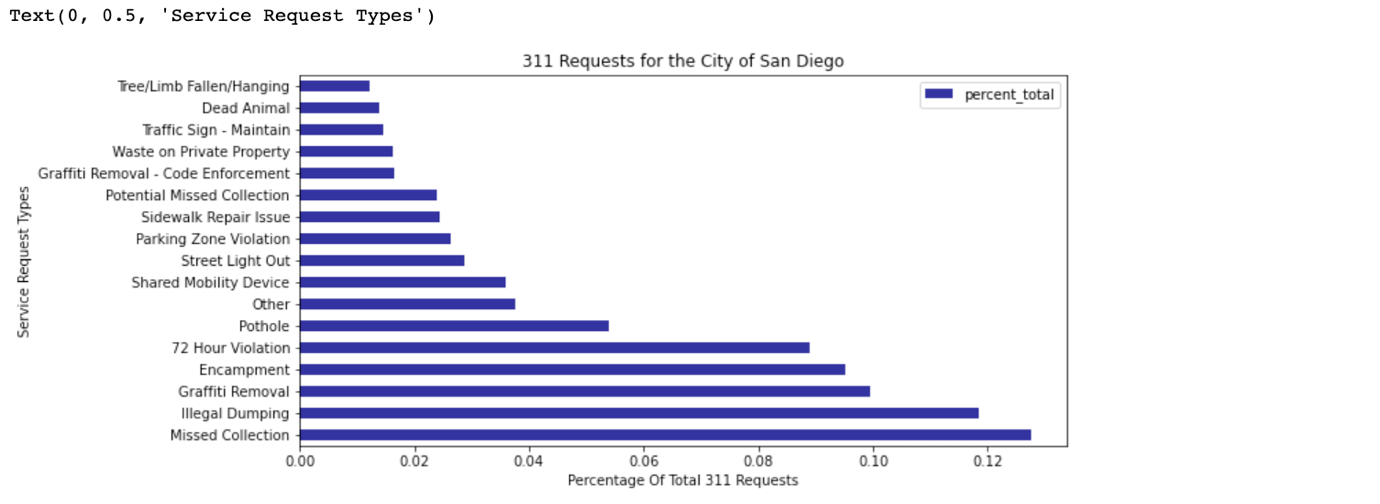
This project is one of the complicated projects in this program. The dataset was obtained from Kaggle. It includes all the data of Airbnb in New York City. The EDA divided the dataset into different areas of NYC. From that, it is clearer to see which area has a higher price and provides better insight to analyze and translate data mining needs to the solution of predicting housing price. After the EDA and data cleaning, the dataset was separated into training and testing set to test the accuracy of R-Part, J48, Naïve Bayes, KNN and Random Forest. From the algorithm evaluation result above, J48 worked best in this project and by using this generated model, the future Airbnb price can be predicted with higher level of accuracy.

IST 736 final project

City service groups should aim to serve their constituents as best they can and cities like San Diego are at the forefront of promoting digital services to reach this objective. The data collected from the Get It Done San Diego app is publicly available and therefore readily available to analyze.

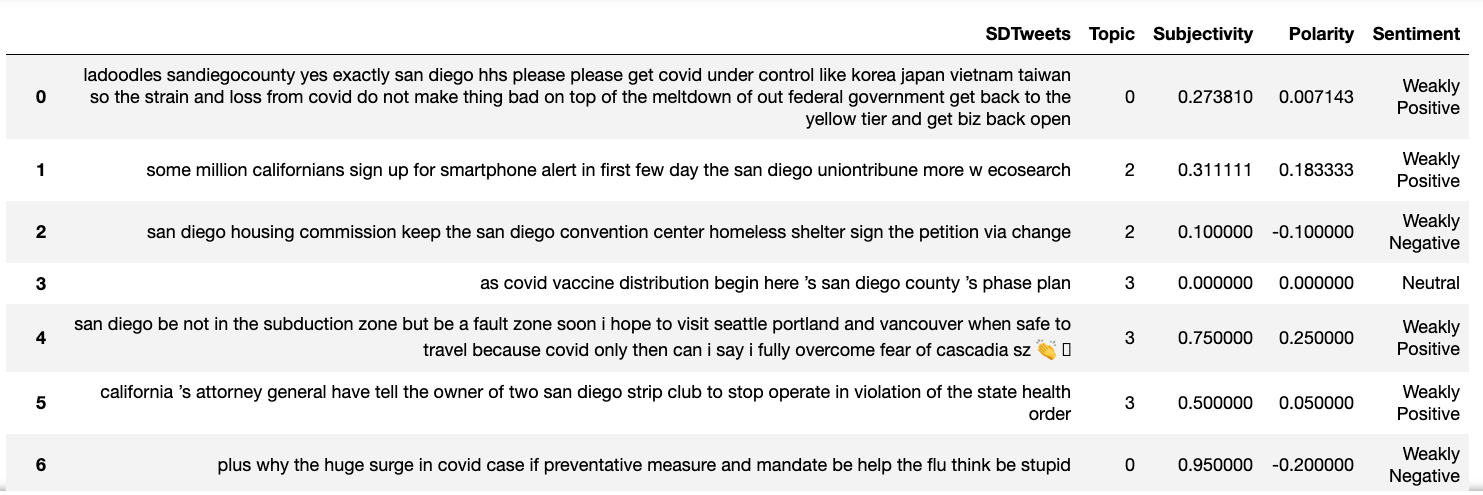
There are or course many other avenues to understand more about the needs of the people of San Diego. Looking at multiple different sources of San Diego related data creates a more complete picture of what is important to the city and its constituents and visitors. Twitter is another great resource to combine with the 311 data. Twitter similarly opens its data to the world (upon approval of a developer account) and is an incredible resource for topic modelling and sentiment analysis.





Sentiment analysis of Tweets

With the three functions applied to the cleaned twitter data the sentiment of the tweets can now be observed.



Based on the nature of service requests, they generally tend to carry a negative sentiment label. Therefore, filtering tweets by negative sentiment and relevant topics allow for a targeted prediction of service requests.

By creating proactive measures for getting ahead of issues before they evolve into bigger ones, city workers can handle them with more efficiency. Social media channels like the Get it Done app and Twitter are great sources of information that can offer immense help in keeping a city functioning. As urban populations continue to grow, so too will the number of non-emergency service requests. Hopefully, these tools will enable San Diego and their counterparts to build up a strong infrastructure for managing them in the future.

The project is unique from all other project provided here in this portfolio. Sentiment analysis is playing major role in text mining. From making text classifications and clustering, I am able to describe basic concepts and methods in text mining. Instead of numbers and figures, text can also be the data to analyze. By implementing sentiment analysis, this project helped me to have a better understanding of the mechanism of advanced text mining algorithms for choosing appropriate technologies for specific text analysis tasks.

Reference

Wu, K. (n.d). IST 659: Database Administration. Retrieved from

https://github.com/Kalsarikannit/MSADS\_Portfolio/blob/master/IST\_659\_Project

Wu, K. (n.d). IST 687: Introduction to Data Science. Retrieved from

https://github.com/Kalsarikannit/MSADS\_Portfolio/blob/master/IST\_659\_Project

Wu, K. (n.d). IST 707: Data Analytics. Retrieved from

https://github.com/Kalsarikannit/MSADS\_Portfolio/blob/master/IST\_659\_Project

Wu, K. (n.d). IST 736: Text Mining. Retrieved from

https://github.com/Kalsarikannit/MSADS\_Portfolio/blob/master/IST\_659\_Project