MS Applied Data Science

Portfolio Milestone

Nianyou Zhang

SUID: 630287321

nzhang01@syr.edu

Syracuse University

School of Information Studies

Table of Contents

[Introduction 3](#_Toc89176518)

[MS Applied Data Science Learning Goals 4](#_Toc89176519)

[Medical Record Database Management System 5](#_Toc89176520)

[Overview: 5](#_Toc89176521)

[Learning goals reflections: 9](#_Toc89176522)

[Topic Modeling and Sentiment Analysis of COVID-19 Vaccine Tweets 10](#_Toc89176523)

[Overview: 10](#_Toc89176524)

[Learning goals reflections: 13](#_Toc89176525)

[Heart Failure Prediction 14](#_Toc89176526)

[Overview: 14](#_Toc89176527)

[Learning goals reflections: 16](#_Toc89176528)

# Introduction

Through January 2020 to December 2021, I studied at Syracuse University School of Information Studies MS Applied Data Science (ADS). During the four semesters, I have taken 12 courses, covering full range of data science life cycle, from data acquisition to storing, data engineering to modeling, and finally model interpretation and obtaining actionable insights.

In this document, I select three of my projects to present. These three projects are all related to medical field but emphasizing on different areas from learning goals:

Medical Database: database management system, focusing on data acquisition, storing and query.

COVID vaccine tweets: focusing on data engineering, analyzing unstructured data.

Hear failure prediction: focusing on model selection and interpretation, analyzing structured data.

I will provide an overview for these three projects, then discuss how they are tied back to the learning goals of our ADS program.

# MS Applied Data Science Learning Goals

The projects presented in this document will demonstrate these goals (program required):

1. Collect, store and access data by identifying and leveraging applicable technologies.

2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.

3. Apply visualization and predictive models to help generate actionable insight.

4. Use programing languages such as R and Python to support the generation of actionable insight.

5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).

6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

# Medical Record Database Management System

IST 659 - Data Administration Concepts and Database Management Spring 2020

## Overview:

Data acquisition and archiving are important steps in data science life cycle. IST 659 database management system (DBMS) course established fundamental understanding of database for me. For the class project, I created this comprehensive entity relation database for medical history.

I consulted Dr. David Larsen from David B. Falk College of Sport and Human Dynamics when I was initially planning to construct a medical history database management system. I learned that in his studies of analyzing healthcare data from low-income countries, the inconsistency and discontinuity of the data is the main obstacle. Essential data are not linked, such as necessary demographic and geographic information, together with medical histories.

My project aims to create a concise, easy-to-implement database management system according to the needs mentioned above, linking demographic and geographic information together with vaccination records and medical history.

The database management system has the following features:

a. Individual central. This will avoid the inconsistency from different hospital records, by tying the records to an individual's citizen ID. This will also create linkage among immunization, diagnosed diseases, symptoms, and treatments, in all the hospitals, to provide access to large set of data for better analysis.

b. The implementation of time frame. This is effective in keeping a consistent, continuous record of an individual. It will spot expired immunization and help emergency brakes in case of epidemics.

c. Treatment assessment. Treatment results are recorded to provide information whether the diseases were cured. This tells the effectiveness of certain treatments and informs the existence of certain antibodies in an individual or proportion of the population.

d. Symptoms-Disease analytics. Symptoms are included for disease diagnosis. As a result, this provides data for disease analytics, assisting doctors with diagnosis.

d. Tracking Cost. All costs are recorded and declared as funded by Federal, State or Private. This will help healthcare assess budgets and individual financial plans.

Here is the entity relation diagram for my database management system:



The actual database management system is created in Microsoft SQL. Here are example tables for vaccination history:

CREATE TABLE vaccineDictionary

(

vaccineId INTEGER NOT NULL,

vaccineName VARCHAR(100) NOT NULL,

vaccineDescription TEXT,

CONSTRAINT pk\_vaccineDictionary PRIMARY KEY (vaccineId),

CONSTRAINT chk\_vaccineName CHECK (vaccineName='Tetanus' OR vaccineName='Hepatitis' OR vaccineName='Human Papillomavirus' OR vaccineName='Mumps' OR vaccineName='Influenza' OR vaccineName='Hib')

);

CREATE TABLE immunizationRecord

(

citizenId INTEGER NOT NULL,

vaccineId INTEGER NOT NULL,

dateAdministered DATETIME DEFAULT GETDATE() NOT NULL,

iClinicName VARCHAR(80) NOT NULL,

immunizationCost DECIMAL NOT NULL,

iFoundingSource VARCHAR(1) NOT NULL,

CONSTRAINT pk\_immunizationRecord PRIMARY KEY (citizenId, vaccineId, dateAdministered),

CONSTRAINT fk\_immunizationRecord1 FOREIGN KEY (citizenId) REFERENCES demographicInformation (citizenId),

CONSTRAINT fk\_immunizationRecord2 FOREIGN KEY (vaccineId) REFERENCES vaccineDictionary (vaccineId),

CONSTRAINT chk\_iFoundingSource CHECK (iFoundingSource='F' OR iFoundingSource='S' OR iFoundingSource='P')

);

Sample data are entered after all the tables are created to test the database. Here is a query example:

-- What are common symptoms for atherosclerotic disease?

SELECT dd.dClinicName ,dd.dateDiagnosed, ddc.diseaseName, sdc.symptomName, sdc.symptomDescription, ddc.diseaseDescription

FROM diagnosedDisease dd

JOIN disease d

ON dd.diseaseId=d.diseaseId

JOIN diseaseSymptom ds

ON d.diseaseId=ds.diseaseId

JOIN diseaseDictionary ddc

ON d.diseaseTypeId=ddc.diseaseTypeId

JOIN symptomDictionary sdc

ON ds.symptomId=sdc.symptomId

WHERE ddc.diseaseName='atherosclerotic disease'

A screenshot of a social media post

Description automatically generated

-- important note, this view protects patient information

CREATE VIEW symptom\_atherosclerotic\_disease AS (

SELECT dd.dClinicName ,dd.dateDiagnosed, ddc.diseaseName, sdc.symptomName, sdc.symptomDescription, ddc.diseaseDescription

FROM diagnosedDisease dd

JOIN disease d

ON dd.diseaseId=d.diseaseId

JOIN diseaseSymptom ds

ON d.diseaseId=ds.diseaseId

JOIN diseaseDictionary ddc

ON d.diseaseTypeId=ddc.diseaseTypeId

JOIN symptomDictionary sdc

ON ds.symptomId=sdc.symptomId

WHERE ddc.diseaseName='atherosclerotic disease'

);

Lastly, a user interface is created via Microsoft Access, here is an example report using UI:

Report question: what are prevalent diseases on Nottingham Rd?

A screenshot of a cell phone

Description automatically generated

## Learning goals reflections:

This project emphasizes the following learning goals:

1. Collect, store and access data by identifying and leveraging applicable technologies.

5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).

6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).

The medical record database management system is created to acquire, store and query relational data. It uses platform Microsoft SQL, Access, Visio.

The medical record database management system is built upon the communication with domain expert, for data needs and data questions.

The medical record database management system is built to protect data privacy: different users can only access data according to their specific needs. I became very aware of the HIPAA law during this project.

# Topic Modeling and Sentiment Analysis of COVID-19 Vaccine Tweets

IST 736 - Text Mining Spring 2021

## Overview:

COVID-19 has been a global disaster since year 2020. Its vaccination was a hot topic during the time I was taking text mining class. Thus, I did my project on analyzing COVID-19 vaccine tweets.

The idea behind this project is to practice analyzing text data and discover tweeter users’ views (both topics and sentiments) toward the COVID vaccine. Hearing completely opposite opinions on COVID vaccine from news and other social medias sparked my interest, so I was hopping to analyze the tweets myself to discover the “ground truth”. Is it true that people’s confidence in COVID vaccine has been steadily going down?

The main tasks of this project are running K-means clustering, LDA topic modeling for data exploration, then Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) for sentiment analysis.

The goal of this project is to analyze text data, thus data engineering is the most essential part of the analysis. The data engineering process is mostly consisted of vectorization and stop words removal. I used TF-IDF vectorizer to penalize extremely common and uncommon words.

I set up a loop to use word cloud to visualize the dataset and tweak the stop words list. Here are examples of word clouds:

**A picture containing text, newspaper

Description automatically generated**

This is a word cloud created by the titles of the tweets.

**A picture containing text, newspaper

Description automatically generated**

This is a word cloud created by the body text of the tweets.

Many swearing words were added to the stop words list through this process.

Here is a comparison of the results of topic modeling without extending the stop words list and after extending the stop words list.

|  |  |
| --- | --- |
| **Topic** | **Key Words** |
| 0 | haha dude joke coincidence excellent stroke flat appreciate gets yes yep immunology misinformation careful breathing really de earth debate wait |
| 1 | welcome dumb explain rules aluminum big deal real science feeling possible deal quack misinformation big bodies downvote sure chracter allergic google kind |
| 2 | like one autism said measles vaccinated read wrong mercury yes cause point still immunity children tell going really anti actually |
| 3 | idiot x200b odd northwest someone pacific northwest pacific trolling f\*\*k bit hope vaxfact site assumed hello posts troll fraudulent bias |
| 4 | lol oh bruh hell her sodium logical shut antivax water says sorry article dihydrogen monoxide joke coat f\*\*king heavy |
| 5 | much mercury something change mean anti vaxxers anti vaxxers compounds elements comments gonna reddit appreciated sub one indeed play stroke compliment |

|  |  |
| --- | --- |
| **Topic** | **Key Words** |
| 0 | **Chemical Components** |
| **mercury sodium elements compounds thimerosal chloride compound** hear brain contains sodium brain damage organic contain indeed elemental consumption food daily toxic |
| 1 | **Unknown** |
| stroke odd assume salt excellent science northwest live risk kids playing pacific controlled approval nurse smith walker children |
| 2 | **Diseases / Medical** |
| like **autism measles** wrong cause read anti really point **vaccinated** saying bad time actually children **immunity** way **immune** someone made |
| 3 | **Agreement** |
| **ok** tell misinformation **welcome** change reason **yes possible** quack unfortunately needle vial spreads lying multi assumed title **loved** allergic ones |
| 4 | **Chemical** |
| day **aluminium ethylmercury** comments questions **mercury dihydrogen monoxide** worry name relevant like appreciated heavy eat bet big deal started immunology |
| 5 | **Unknown** |
| yes joke dumb coincidence food compliment vaccinated youre work rules thought works antivaxxers god damn myth consequences feeling kid bias |

This topic modeling step helped me understand people’s common concerns of COVID-19 vaccine.

For sentiment analysis however, the stop words were all kept because many stop words can express sentiments.

I run into problems when doing sentiment analysis: the dataset was unsupervised, and I cannot train models without labels. Thus, I developed two strategies to approach this task: 1. Train models with similar dataset that has sentiment label; 2. Use pretrained models such as TextBlob.

For the first approach, I used two different algorithms to build sentiment prediction models using a similar dataset: Multinomial Naïve Bayes(MNB), Support Vector Machines(SVM). The result was quite disappointing mostly because the “similar” dataset I found was for drug review and it is quite different from COVID-19 vaccine discussion tweets.

For the second approach, I used TextBlob to classify the tweets, then hand labeled tweets to evaluate the result. I discovered that most of the tweets are neutral.

## Learning goals reflections:

At the end of the project, I have learned people’s common concern about the COVID-19 vaccine. Some people worried about the harmful ingredient of the COVID-19 vaccine and others worried about the side effects (risk of getting diseases) from the vaccinations.

This project emphasizes the following learning goals:

2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.

3. Apply visualization and predictive models to help generate actionable insight.

4. Use programing languages such as R and Python to support the generation of actionable insight.

# Heart Failure Prediction

IST 707 - Applied Machine Learning Fall 2020

## Overview:

Modeling is the core step of analyzing data. Parameter searching and model selection requires both technical and domain expertise. During my study of IST 707 Applied Machine Learning, I applied the strategies and knowledge of machine learning models in a task of predicting heart failure event with patients’ record of pre-existing conditions and blood-test results.

The purpose of this project is to build predictive models with different algorithms and compare their performance. The prediction task is to decide weather a patient will experience heart failure. The burning question for this project is to discover what preconditions and blood-test results are related to heart failure; are the current measurements enough to train models to predict heart failure.

The measurements collected in this dataset are as following:

|  |  |  |
| --- | --- | --- |
| Basic information: | | |
| Attribute name | age | sex |
| Data type | num | num, binary |
| Description | age of the patient | 0 for female, 1 for male |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pre-existing conditions: | | | | |
| Attribute name | smoking | diabetes | high\_blood\_pressure | anaemia |
| Data type | Boolean | Boolean | Boolean | Boolean |
| Description | if the patient smokes or not | if the patient has diabetes | if the patient has hypertension | decrease of red blood cells or hemoglobin |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Blood-test results: | | | | | |
| Attribute name | platelets | diabetes | high\_blood\_pressure | anaemia | ejection\_fraction |
| Data type | num | num | num | num | num, percentage |
| Description | platelets in the blood (kiloplatelets/mL) | level of the CPK enzyme in the blood (mcg/L) | level of serum creatinine in the blood (mg/dL) | level of serum sodium in the blood (mEq/L) | percentage of blood leaving the heart at each contraction |

I firstly did association rule mining to discover what preconditions and blood-test results are closely related to heart failure. Here is the result of association rules:

After a trying out a range of support and confidence values, the following combination yield most appropriate number of rules.

* Support=0.05;
* Confidence=0.7;
* Minlen=4;

This combination results in 10 rules. The best rule is:

|  |  |  |
| --- | --- | --- |
| Rule: | LHS | RHS |
| ejection\_fraction=1st | DEATH\_EVENT=Yes |
| serum\_creatinine=4th |
| serum\_sodium=1st |
| Parameters: | Support | 0.05 |
| Confidence | 0.79 |
| Lift | 2.46 |

The confidence of 0.79 is quite strong, because 79% patients with such attributes passed away. The lift of 2.46 is not extremely high, but high enough to indicate the dependence between the three conditions and death.

After association rules, I built a predictive model to predict heart failure event. During this task, I tested three classic machine learning models: support vector machine (SVM), decision tree and artificial neural network (ANN).

For SVM, I performed grid search with a range of parameters to obtain the best results under each different boundary type:

|  |  |  |  |
| --- | --- | --- | --- |
| Type: | SVM linear | SVM polynomial | SVM radial |
| Test Accuracy | 0.8305 | 0.7288 | 0.7458 |
| Train Accuracy | 0.7625 | 0.7792 | 0.8292 |
| Specificity | 0.8462 | 0.7143 | 0.6667 |
| Precision | 0.95 | 0.95 | 0.9 |
| Recall | 0.7755 | 0.8837 | 0.8182 |
| F\_measure | 0.8539 | 0.9157 | 0.8571 |

SVM polynomial has highest F\_measure and recall, thus become the choice.

After running other models, I compared ANN, SVM polynomial and decision tree together:

|  |  |  |  |
| --- | --- | --- | --- |
| Type: | ANN | SVM polynomial | DT |
| Test Accuracy | 0.8136 | 0.7288 | 0.6441 |
| Train Accuracy | 0.7542 | 0.7792 | 0.7792 |
| Specificity | 0.7857 | 0.7143 | 0.3333 |
| Precision | 0.925 | 0.95 | 0.9 |
| Recall | 0.7708 | 0.8837 | 0.9474 |
| F\_measure | 0.8409 | 0.9157 | 0.9231 |

It was a hard choice between ANN and SVM polynomial. For this dataset, model accuracy and recall are both important to consider. Recall is important because in diagnosing patients, it is safer to have false positive than false negative. From above evaluation measures, ANN has best testing accuracy, and good overall measures. When I take in consideration of training set accuracy, ANN is much less biased than SVM. The conclusion for predictive model is: Artificial Neural Network is a best choice.

## Learning goals reflections:

From my project, I learned that blood test results are strong indicators of fatal heart conditions. Patients with decrease of red blood cells or hemoglobin have nearly 9% increased heart failure risk. Blood with high serum creatinine and low serum sodium are closely associated with heart failure. An actionable insight from this conclusion is that for people that have concern over their heart condition, they should take blood test regularly to monitor the key parameters mentioned above.

This project emphasizes the following learning goals:

2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle.

3. Apply visualization and predictive models to help generate actionable insight.

4. Use programing languages such as R and Python to support the generation of actionable insight.

6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy).