Syracuse University, School of Information Studies
M.S in Applied Data Science

# Portfolio Milestone

Yibo Feng
SUID: 57533926
https://github.com/yfeng0308/MSADS_Portfolio

Table of Contents

1. Introduction ..... 3
2. IST 659: Database Administration ..... 3
Project Description ..... 3
Reflection & Learning Goals ..... 5
3. IST 664: Natural Language Processing ..... 6
Project Description ..... 6
Reflection & Learning Goals ..... 7
4. IST 707: Data Analytics ..... 8
Project Description ..... 8
Reflection & Learning Goals ..... 10
5. IST 718: Big Data Analytics ..... 11
Project Description ..... 11
Reflection & Learning Goals ..... 13
6. Conclusion ..... 13
7. References ..... 15

## 1. Introduction

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 various tools and techniques. Courses like Database Administration (IST659), Natural Language Processing (IST664), Data Analytics (IST707) and Big Data Analytics (IST718). Project reports and presentations deliver insight using tools such as SQL Server Management Studio, R Studio, Spyder. During the time of program, these courses also give me opportunities to be familiar with popular programming language such as Python and data analyzing language like R. The Applied Data Science program has seven learning objectives, and these will be exemplified by the project review in this portfolio:

- Describe a broad overview of the major practice areas in data science.
- Collect and organize data.
- Identify patterns in data via visualization, 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 analyzes.
- Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
- Synthesize the ethical dimensions of data science practice.

## 2. IST 659: Database Administration

### Project Description

This project focuses on designing a database for Syracuse University iSchool course catalog and course registration system. The iSchool is offering a course catalog which entails available classes of different levels pertinent to various information areas. Currently, the course catalog can be accessed through 2 separate routes which simply display a list of available courses and their basic information (course code, instructor, overview, and pre-requisite course). Apart from the course catalog, the course registration system is integrated with the class searching system called 'Myslice' where users can check their current credit records and view available courses. In conclusion, the current course catalog and course registration system are separate,

and they are not capable of providing one single platform for comprehensive data sources, resulting in user inconveniences. Therefore, the purpose of this system development is to establish an integrated iSchool course catalog database and link it to the course registration system for enhancing user conveniences. For instance, students are expected to shorten their time for browsing and selecting courses they are interested in or required to take and make better course registration quality-wise. Also, administrators can analyze such data more conveniently and make use of the analysis for improving iSchool academic course design.

Logical model and ER diagram were developed to organize the relationships between Job, JobOffered, Company, ClassRoom, Instructor, Class, TechTool, Course, Student, Registration. Tables were created in SQL Server Management Studio and data population was accomplished by Microsoft Access. As Figure 1 shows that, this is the visualization network of ten tables. Building several forms to display multiple information. One is displaying details of classes for a specific course. One is displaying available job information for students.

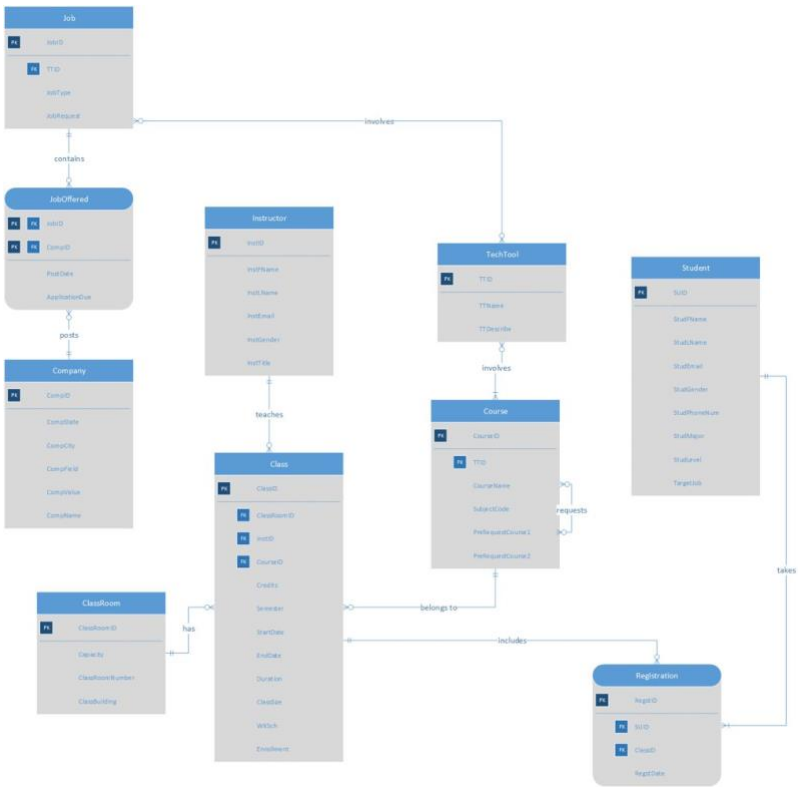


Figure 1: Logical Model/ER Diagram

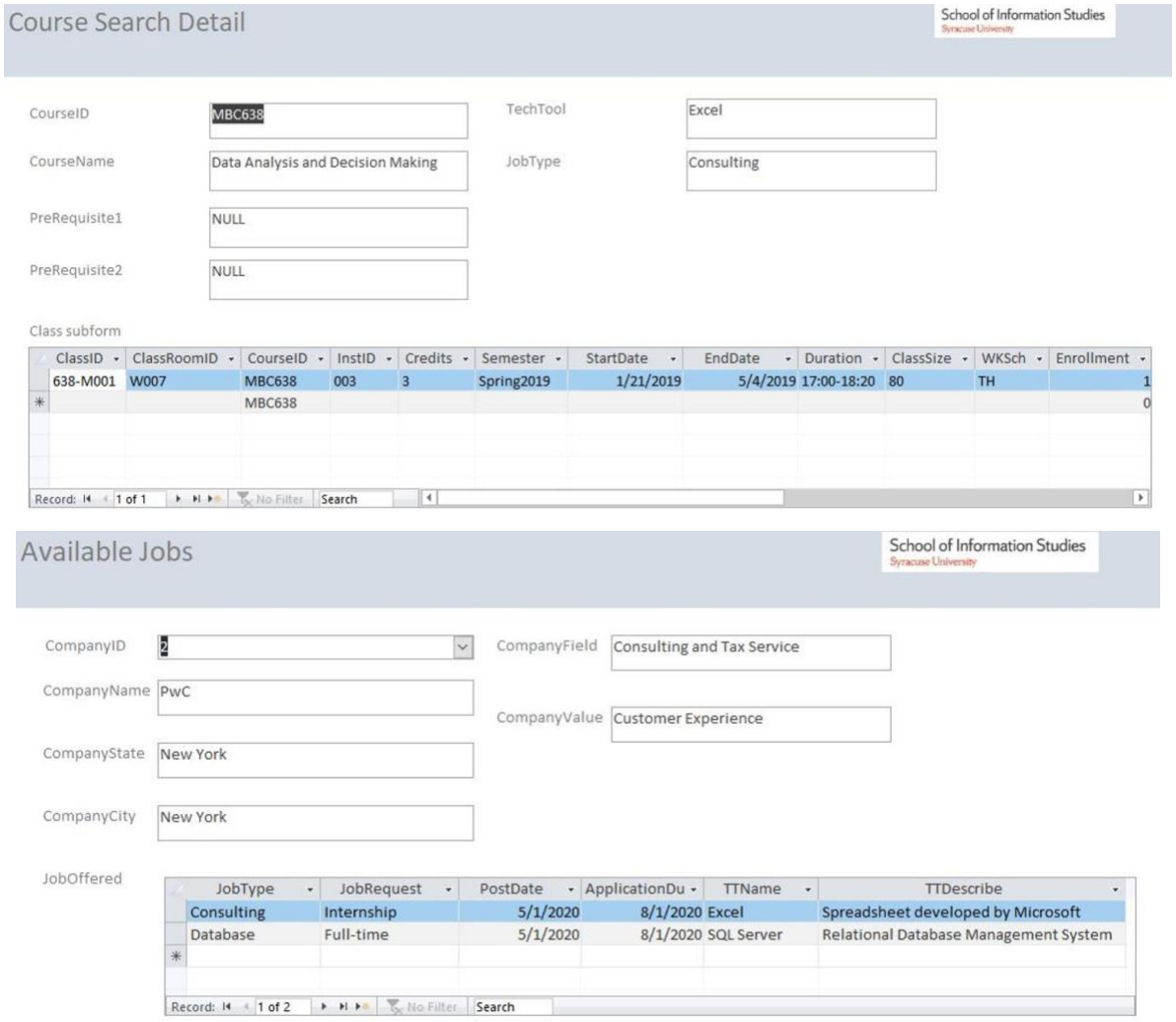


Figure 2: Form Examples

**Reflection & Learning Goals**

This course provides opportunity to develop a data management solution which reveals the importance of data is stored and accessed. This project contributed to the ability to deliver actionable insight to the field of program management. This database help students to view their profile as well as credit status and help instructors notify course updates and companies could use this system to find potential employees by querying it. This is also significant for data analysts and data scientists. Overall, this project contributed successful application of learning goals such as collect and organize the data and identify the pattern via visualization. Thinking of three perspective point (Student, Instructor and Company) reveals there is a plan of action to implement business decisions.

## 3. IST 664: Natural Language Processing

### Project Description

With more and more text data generated by the Internet, the problem of text information overload is becoming more and more serious. Text summaries are designed to transform text or text collections into short summaries containing key information. According to the output type, it can be divided into extractive abstract and generative abstract. Extractive abstract extracts keywords from the source document to form an abstract. Generative abstracts generate new words and phrases based on the original text to form an abstract. Broadly speaking, there are two ways to summarize the text: Extraction-based summarization and Abstraction-based summarization. This project is to use extraction-based summarization technique based on two algorithms: TF-IDF and TextRank. The goal is to figure out which algorithm is suitable to summarization of News.

Based on article "Text Summarization Techniques: A Brief Survey", text summarization got attraction in the 1950s. There was a significant research which explained a method to extract sentences from text or documents using features like word or phrase frequency. TFIDF is such an algorithm mainly using word frequency to measure sentences' weights and extract sentences. The reason we choose TextRank because the core of it is to build similar matrix. Similarity of two sentences in similar matrix is calculated by computing cosine similarity with TFIDF weights for words. It could conclude that two algorithms have connections but run with different tracks. Therefore, our goal using them is to see how different result they could bring. Figure 3 is specific process for TF-IDF and Figure 4 is for TextRank.

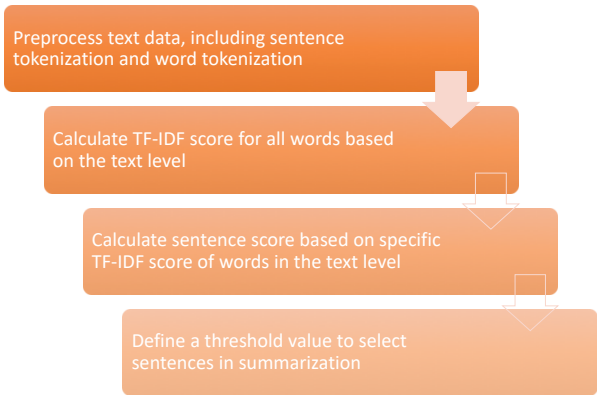

Figure 3: Text Summarization Process in TFIDF

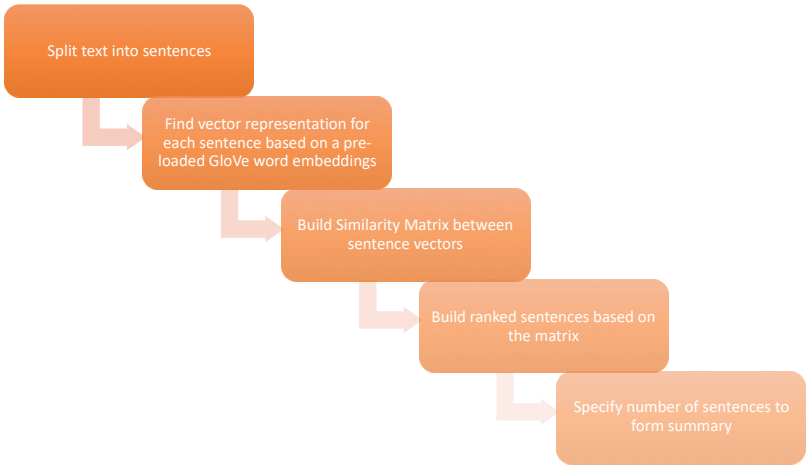


Figure 4: Text Summarization Process for TextRank

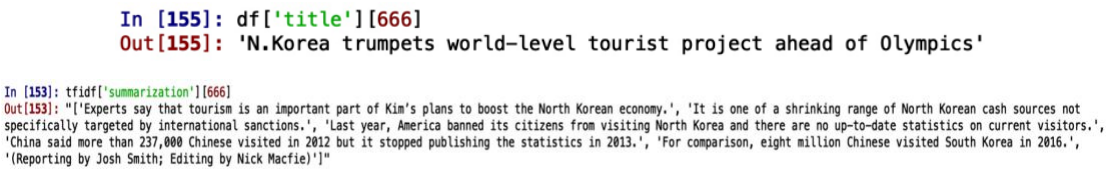


Figure 5: Example Summarization for TFIDF


Figure 6: Example Summarization for TextRank

**Reflection & Learning Goals**

Based on Figure 5 and Figure 6, It is clear that summarization for TextRank extract summarization is more accurate than it for TFIDF. The reason is that the first sentence of summarization for TextRank shows key word such as North Korea, World-Level and tourist project. However, it is hard to extract these words if we see the first several sentences from summarization for TFIDF. Also, I have to say there exist limitation for these two algorithms and

dataset. It is because it is hard to use existing columns from dataset to evaluate the accuracy of all summarization based on extraction summarization.

Text summarization is a common problem in natural language processing. In this project, we preprocess dataset by sentence tokenization, removing punctuation as well as number and special characters, using all lowercase and removing stopwords. and use ROUGH to evaluate two models' performance. Compared the results of two model, the TextRank has a better performance than TF-IDF.

Overall, this project contributed successful application of learning goals such as collect and organize the data. Although all collected data is public record, consideration must be made to ensure that only the relevant information is requested to both balance request limitations and user privacy.

## 4. IST 707: Data Analytics

### Project Description

Through studying Data Analytics under the direction of Prof. Jesse Cases, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering. The project dataset I chose collected top products information including ratings and sale performance from E-commerce platform "Wish". At the beginning of data mining, I started to work on product segmentation. The goal is to help the platform improve their customer base, work on target areas, and segment customers based on purchase history and interests. Besides, product segmentation will help the platform find out the common points and different types of product, then they will be able to set target sale policy. I was also interested in what factors contribute to high rates of rating with five count which means the percentage of rating 5 in total rating count is higher than 50%. This will contribute to improving customer preference. I also want to analyze potential factors that help increase the sale of products. For example, it is interesting whether the merchant rating will influence the sale of products. By analyzing these factors, the platform will be able to predict sales status of each product.

This data required the cleaning and preprocessing of dataset which included removing irrelevant columns and reformatting category feature to numeric ones. This project included

analysis from The project included analysis from three directions. The first one was product segmentation. The data mining task is cluster analysis by using K-Means. From the left side of Figure 7, I chose 4 as elbow point and put it into K-Means model. The right side of Figure 7 was clearly showing that the clusters were not clear and there were a lot of outliers. Then, we chose another model hierarchical cluster and used Euclidean distance to build the distance matrix. The final accuracy for this model is 0.998.

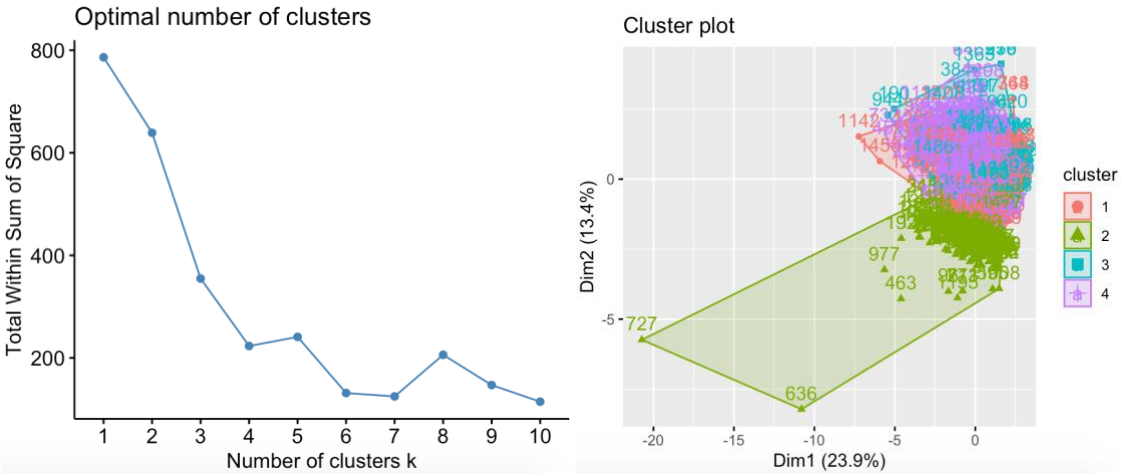


Figure 7: Cluster Model using K-Means

The second direction was to find factors for High Rating Rates. The specific preprocessing was to reformat category variable as factors and convert numeric variables into category based on their 1st and 3rd quartiles. The data mining task used in this part is Association Rule. Based on following rule from Figure 8, it could conclude that products with badges of product quality would contribute to gaining high rate of 5 ratings. Badges of fast shipping and local product wouldn't influence the high percentages of the rating 5. Origin country of China would contribute to high level of rating.

| lhs | rhs | support | confidence | lift |
|---|---|---|---|---|
| {badge_product_quality=1, badge_fast_shipping=0, shipping_is_express=0} | {rating_level=high} | 0.07138092 | 1 | 2.653097 |
| {badge_local_product=0, badge_product_quality=1, badge_fast_shipping=0, shipping_is_express=0} | {rating_level=high} | 0.06871247 | 1 | 2.653097 |

| | | | | |
|---|---|---|---|---|
| {badge_local_product=0, badge_product_quality=1, badge_fast_shipping=0, shipping_is_express=0, origin_country=CN} | {rating_level=high} | 0.06604403 | 1 | 2.653097 |
| {badge_product_quality=1, badge_fast_shipping=0, shipping_option_price=<3} | {rating_level=high} | 0.06204136 | 1 | 2.653097 |

Figure 8: Association Rule Result

The final direction was to make sale status prediction. During preprocessing, I selected variables that might contribute to classifying high sales and set high 'units_sold_level' as 1 while others as 0. The main data mining task was to implement classification algorithm such as SVM and ANN. In this part, four model including KNN, Random Forest, SVM and ANN would be built to compare result with each other. Figure 9 provided result from each model. Since the training accuracy and prediction accuracy of Random Forest differed a lot, there may exist overfitting in the model. Besides, based on dataset description I found the positive value is 0, and what met the goal was to classify unit_sold_level=1. Therefore, recall and F-measure could be better to evaluate the models. According to this, SVM will be a better model for this dataset.

| Model | Training Accuracy | Prediction Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| KNN | 0.6974428 | 0.6756 | 0.6973 | 0.9100 | 0.78957631 |
| Random Forest | 0.7958761 | 0.7023 | 0.9744 | 0.5700 | 0.71925408 |
| SVM | 0.6691738 | 0.6689 | 0.6689 | 1 | 0.80160585 |
| ANN | 0.6791207 | 0.6622 | 0.7071 | 0.8450 | 0.76992397 |

Figure 9: Classification Result

**Reflection & Learning Goals**

From the first direction, I found the unexpected result compared with class example using cluster model. It illustrated it is important to test different data mining techniques to find out the simplest, most accurate prediction models. Using alternative strategies and weighting the benefit

of each technique with combination of characteristics of dataset could make the whole process more effective and convincible and provide higher precision in data mining tasks. This project contributed to successful application of the learning goal by using alternative strategies based on data and demonstrate communication skills regarding data and its analysis. Combination of data mining and visualization to identify patterns in the data were used in classification tasks.

## 5. IST 718: Big Data Analytics

### Project Description

The main purpose of the project was to figure out how the clothes brand, Rent the Runway, should focus on utilizing the factors that could affect customer feedback and predicting which kinds of clothes should be promoted to a unique customer. To complete that object, the first thing I did was to extract the columns that can illustrate customer feedback. Then based on different data types of those columns, targeted data processing had been given. For example, a sentiment analysis was made to cope with column "review_text" and column "review_summary" which contain string data type. To predict customer's rating feedback for a special product (numeric data type), I used linear regression algorithm to run a machine learning.
There were mainly four evaluations in this project. They were Sentiment Analysis, Customer Rating Analysis, Recommendation Analysis and Product with High Rating Analysis. There were three methods using to analyze four evaluations such as sentiment analysis, clustering analysis and classification analysis.

Figure 10 shows 10 most negative words and Figure 11 shows 10 most positive words through sentiment analysis. We were focused on creating an ideal sentiment analysis model which was used to predict the sentiment of reviews in each transaction for the purposes of finding out important factors influencing customers attitudes. The model followed Logistic Regression. Having added the new column "Attitude", we randomly split the dataset into 3 parts: training (60%), validation (30%) and testing (10%) for this part's analysis. After a series of text processing on "review_text" column, we finally got the output column "tf-idf". Then we set the feature column "tf-idf" and the label column "Attitude" to the two logistic regression models.

| | word | score |
|---|---|---|
| 0 | unflattering | -0.287943 |
| 1 | disappointed | -0.266987 |
| 2 | awful | -0.224330 |
| 3 | ridiculous | -0.208293 |
| 4 | disappointing | -0.205993 |
| 5 | cheap | -0.182909 |
| 6 | returned | -0.181447 |
| 7 | linebacker | -0.180583 |
| 8 | unfortunately | -0.174893 |
| 9 | way | -0.174409 |

Figure 10: 10 Most Negative Words

| | word | score |
|---|---|---|
| 0 | compliments | 0.418912 |
| 1 | comfortable | 0.351355 |
| 2 | perfect | 0.324392 |
| 3 | wore | 0.279257 |
| 4 | loved | 0.232826 |
| 5 | great | 0.213506 |
| 6 | little | 0.204051 |
| 7 | true | 0.105880 |
| 8 | perfectly | 0.104809 |
| 9 | bit | 0.102677 |

Figure 11: 10 Most Positive Words

During analysis on Product with High Rating, the goal I implemented this model was to predict the situation where customers are likely to give different ratings for the purpose of judging whether the customer would like to provide a high rating or not. The model using in task was Random Forest classification. I split the transformed dataset into three parts including training(60%), test(30%) and validation(10%). After parameter modification and model operation, I extracted all features and sorted them by feature importance in descending order. From Figure 12, “fit”, “rented_for” and “age” are top 3 most significant features in this task. I didn’t find an efficient way to visualize the result of random forest model and therefore I chose one tree example showed in Figure 13.

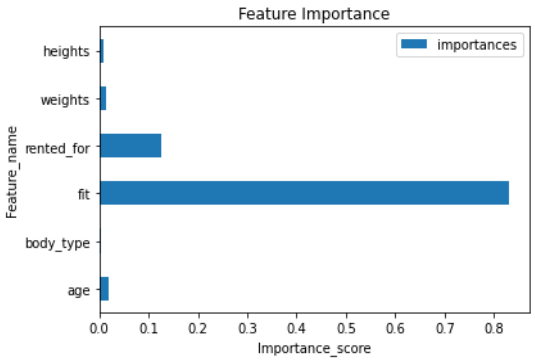


Figure 12: Feature Importance

```
Tree 17 (weight 1.0):
  If (feature 2 in {1.0})
   If (feature 0 <= 29.5)
    Predict: 1.0
   Else (feature 0 > 29.5)
    If (feature 1 in {0.0,1.0,4.0,6.0})
     If (feature 5 <= 189.23000000000002)
      Predict: 1.0
     Else (feature 5 > 189.23000000000002)
      If (feature 3 in {1.0,2.0})
       Predict: 1.0
      Else (feature 3 not in {1.0,2.0})
       Predict: 0.0
    Else (feature 1 not in {0.0,1.0,4.0,6.0})
     Predict: 1.0
  Else (feature 2 not in {1.0})
   If (feature 4 <= 111.5)
    If (feature 3 in {1.0,6.0})
     If (feature 5 <= 171.45)
      Predict: 1.0
     Else (feature 5 > 171.45)
      If (feature 2 in {0.0})
       Predict: 2.0
      Else (feature 2 not in {0.0})
       Predict: 1.0
    Else (feature 3 not in {1.0,6.0})
     If (feature 5 <= 173.99)
      Predict: 1.0
     Else (feature 5 > 173.99)
      Predict: 0.0
   Else (feature 4 > 111.5)
    Predict: 1.0
```

Figure 13: Tree Example

As for Recommendation Analysis, the goal I implemented this model was to Predict what kinds of product will fit different customers for the purpose of recommending appropriate product to customer. The model using in this task was K-Means Clustering. Since it was a clustering problem, I used “ClusteringEvaluator” and evaluator metric is silhouette score. Based on Figure 14, I set K from 2 to 3 and found the highest silhouette score when k was 3. I then built a cluster result plot based on K-means model. From Figure 15, the cluster result was not bad, and some points were a little bit far away from their clusters.

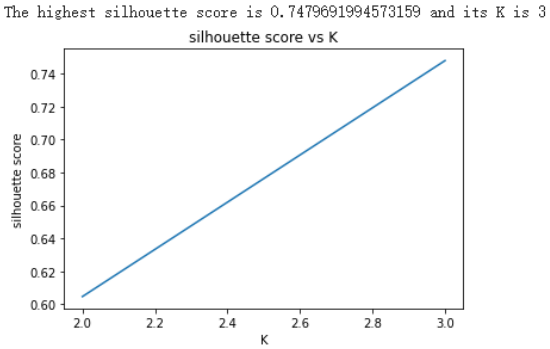


Figure 14: Silhouette Score Plot

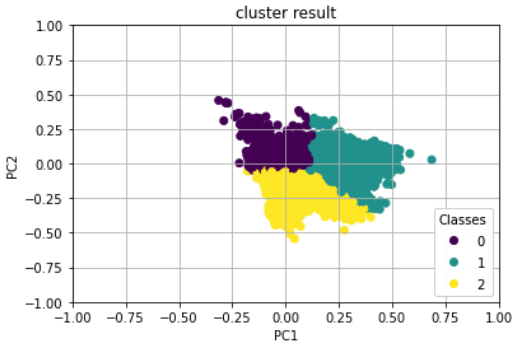


Figure 15: Cluster Result Plot

**Reflection & Learning Goals**

This project provided the opportunity to organize and analyze sale information using data mining techniques, as well as visualization to identify patterns for. It was also necessary to develop a plan of action to quantify the insights developed in this analysis, which translates to measurable and actionable recommendations. Ethical considerations were also necessary to ensure that customer segmentation and profiling was free of bias, using age and fit to profile the previous behavior of a customer, rather than using said information to explain their behavior. This project allowed the data to guide the analysis, requiring alternative strategies to be developed as observations were made within the data.

## 6. Conclusion

This portfolio demonstrates the implementation of learning objectives and practices in the field of data science. All data were collected and organized using online data resources and

programming skills with combination of database administration to be analyzed using statistics and data mining techniques such as clusters analysis and classification analysis from IST 659, IST 707 and IST 718. Data visualization were using to identify patterns which directed to respective analysis. Multiple recommendations were developed to reflect business decisions. For instance, in the project of IST 718, using analysis result of product with high rating to figure out customers with different preference, giving sales department advice to guide them to provide different advices and improve sales performance. Communication skills were developed and displayed during project presentations. Expressing them in the terms which could simply understand. The ethical dimension of data science practices was also reinforced in the application by selecting only relevant data columns and consider user privacy when analyzing personally unique information. These projects were representatives of execution of the learning objectives and have used the skills in the field of data science.

## 7. References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., D., E., B., J., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Applications, 8(10). doi:10.14569/ijacsa.2017.081052

Brownlee, J. (2020, August 21). How to Use ROC Curves and Precision-Recall Curves for Classification in Python. Retrieved November 28, 2020, from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/.

Misra, R., Wan, M., & McAuley, J. (2018, September). Decomposing fit semantics for product size recommendation in metric spaces. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 422-426).

Panchal, A. (2020, November 29). Text Summarization using TF-IDF. Retrieved November 30, 2020, from https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3