# M.S Applied Data Science

# Portfolio Draft

## Submitted by

## Jaya Varshini Prabakar

## SUID: 463948991

### NetId: [jprabaka@syr.edu](mailto:jprabaka@syr.edu)

### Github Link: <https://github.com/jayavarshini6/IST-782-Applied-Data-Science-Portfolio>

### Table of Contents:

|  |  |  |
| --- | --- | --- |
| **S.No** | **Content** | **Page No** |
| 1. | Introduction | 3 |
| 2. | Academic Learning Outcomes | 4 |
| 3. | Project 1 – Healthcare Cost Information Analysis | 4 |
| 3.a | Project Description | 5 |
| 3.b | Project Learning Outcomes | 20 |
| 3. | Project 2 – FudgeInc Data Warehouse Implementation | 20 |
| 3.a | Project Description | 20 |
| 3.b | Project Learning Outcomes | 22 |
| 4. | Project 2 – Los Angeles Crime Dashboard | 22 |
| 4.a | Project Description | 23 |
| 4.b | Project Learning Outcomes | 24 |
| 5 | Project 3 Lacrosse Live Datapipeline Implementation | 25 |
| 5.a | Project Description | 26 |
| 5.b | Project Learning Outcomes | 28 |
| 6. | Conclusion | 29 |

### INTRODUCTION:

The ever-evolving landscape of today's digital age, where data reigns supreme, its influence permeates every facet of our lives, from healthcare to electronics, universities to farming. Within this intricate tapestry, Data Science emerges not merely as a discipline but as a multifaceted spectrum, encompassing the entire lifecycle of data. This dynamic field presents an enticing panorama of boundless learning opportunities, enticing those with a thirst for knowledge and a drive for continuous growth. As one consumed by an insatiable curiosity, drawn to the allure of uncovering hidden insights and unraveling the mysteries held within data, I found myself inexorably pulled towards the pursuit of a Master’s degree in Applied Data Science.

Before embarking on this transformative journey, I found myself immersed in the realm of data as a Data Engineer, where I laid the groundwork for understanding the intricacies of data extraction, storage, and visualization. This foundational experience sparked a fervent curiosity within me, propelling me towards further exploration and ultimately driving me to pursue advanced studies in the field. Throughout my two-year program, I not only delved deeper into the principles of Data Science but also had the privilege of applying my newfound knowledge in real-world settings.

As a Tableau Developer at the Institute of Veterans and Military Families, I endeavoured to translate raw data into accessible visualizations, ensuring that individuals of varying levels of data literacy could glean valuable insights. Moreover, I made it a priority to ensure accessibility for those with disabilities, recognizing the importance of inclusivity in data-driven decision-making processes.

With a solid foundation in Data Engineering, I honed my focus on Data Pipeline and Platform development, complemented by a deep dive into Visual Data Analytics. Leveraging the flexible coursework offered by the iSchool, I tailored my secondary track to align with my unique interests and aspirations. Through these projects, I not only mastered core concepts but also cultivated essential skills in team collaboration, delegation, and the strategic utilization of data to drive impactful decision-making.

This portfolio serves as a testament to my journey, encapsulating the essence of my experiences and highlighting the transformative impact of my top three projects.

### Academic Learning Outcomes:

The essential aspect of the Graduate Applied Data Science program is its Learning outcomes in each of the courses. Every course work that I have taken so far has had 1 or more of these mentioned learning outcomes, which are citated from the ischool’s Applied Data Science official website.

* Collect, store, and access data by identifying and leveraging applicable technologies
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
* Apply visualization and predictive models to help generate actionable insight
* Use programming languages such as R and Python to support the generation of actionable insight
* Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
* Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

### Projects:

#### Healthcare Cost Information Analysis

**Course:** IST 687 Introduction to Data Science

**Project Description:**

A health maintenance organization (HMO) is a network or organization that provides health insurance coverage for a monthly or annual fee. An HMO is made up of a group of medical insurance providers that limit coverage to medical care provided through doctors and other providers who are under contract with the HMO.

This project aims to analyze the HMO dataset to identify types of candidates who are more expensive, and why. The main goals are to use data analysis techniques to predict people who are likely to spend more money on health in the next year and provide actionable insights on how their healthcare costs can be reduced. A review and analysis of healthcare cost data can be a helpful tool for understanding the most common and expensive health conditions where claims have been made; examining trends in costs over time; and comparing utilization rates to local, state, or national norms. Analysis of trends in healthcare expenditures will assist with assessing the effects of health promotion programs.

**Dataset Description** :

The dataset we use consists of various patients' healthcare information in an HMO. It consists of 7582 patient records and 14 columns.

**Variables used:**

X: index of observation

age: Age of the person

location: State of residence

location\_type: type of residence (urban or country)

exercise: exercise status Smoker: if the person smoked during the last year.

bmi: body mass index of the person

yearly\_physical: if the person had a yearly check up with their doctor

Hypertension: Describes if the person had Hypertension

Gender: Type: Categorical

Education\_level: level of college education the client received

Married: marital status

Num\_children: number of children the individual has

Cost: cost of healthcare for the client in the past year

For our initial exploratory Data Analysis, we used all the variables

**DATA ANALYSIS**

Descriptive Summary Firstly, we generated a basic descriptive summary for all the variables.

**EDA:**

**A screenshot of a computer screen

Description automatically generated**

**Box plot for Outlier Analysis:**

**A graph of a function

Description automatically generated with medium confidence**

From the histogram of cost, we see a very right skewed distribution. Telling from the plot, most cases have the cost smaller than 20,000 dollars. Engaging the information from the descriptive summary, the middle 50 percent of cost goes from 970 dollars to 4775 dollars, and maximum cost is 55,715 dollars, which indicates that an outlier detection is very necessary.

**Histogram for distribution analysis**

From the above generated histograms, we can see that the red coloured bars represent the non- expensive clients and the blue bars represent the expensive clients. Both the red and blue bars follows uniform distribution and the red bar is more concentrated and have major peaking at 26 - 28 BMI index which means that more number of non-expensive clients are Overweight and when looking at the blue bar , it is majorly concentrated and has peaking at 34-36 BMI index which means that more number of expensive clients are having the problem of Obesity.

**A graph of a number of bmi

Description automatically generated**

**Correlation analysis**

**Correlation Between numerical Variables**:

For the correlation , We have used the Spearman method for age , bmi , children , cost variables because the data of these variables are which are plotted are not symmetrically distributed and there is skewness present in this data of these variables. The correlation index ranges between -1 to +1 and +1 -represents strong positive correlation . -1 -represents strong negative correlation. Before diving into detailed correlation using scatterplot, a correlation matrix is used.

**A blue and white squares with red text

Description automatically generated**

**Map for cost distribution:**

**A map of the united states

Description automatically generated**

The dataset covers only the North eastern parts of the United States. From the dataset the average cost in New York State is relatively low compared to the other states like Pennsylvania, Connecticut, Massachusetts.

**Data partition for training and test:**

With all the information from descriptive statistics, we will now use it to work on our predictive models and we will be using expensiveness as the dependent variable. The models with be fed by 70 percent of overall observations which work as training dataset, and the model will be tested with the rest 30 percent for predictive accuracy.

**PREDICTIVE MODELS:**

**Linear regression model :**

Initially, we performed linear regression to get a general idea of the data and determine the relationship between the dependent and independent variables. Based on the results of Linear regression, the variables Age, BMI, Children, Smoker, Exercise and Hypertension are the most important variables that have a high effect on determining the amount of money an individual spends on healthcare.

**A computer screen shot of a computer code

Description automatically generated**

When we observe the median which is -352 which tells that it is not symmetrically distributed and it shows some skewness.

The standard errors around the estimates of slope and intercept show the estimated spread of the sampling distribution around these point estimates.

The adjusted R -Squared is 57 % which means that the Age, BMI, Children, Smoker, Exercise and Hypertension, gender (predictors) accounts to 57% variability of cost variable (independent variable). When talking about the major B-Weights contributors - Exercise , Location(Maryland) , Hypertension , Age , Bmi , Children played the major role in creating the effect on the cost variable.

**Logit Regression**

This algorithm is generally a statistical model that models the probability of the event taking place by having log-odds for the event be a linear combination of one or more independent variables. It is based on the inverse logit function. This Logit function comes from the family of Generalized linear models and it is very helpful in predicting Categorical values such as TRUE or False , 0s or 1s ( Binary Outputs). We got an accuracy of about 83% which is good in predicting the Expensive variable and tells us that the Age, BMI, Children, Smoker, Exercise and Hypertension, gender helps in change of variability of expensive variable.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Neural networks**

Neural networks, also known as artificial neural networks, are a part of machine learning and have gradually become very important in deep learning algorithms. A neural network model is composed of an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an according weight and threshold. With its special characteristic of hidden layers, the neural network model provides more flexibility and profoundness in exploring the data pattern. In this HMO data consulting case, we are going to put training data into a neural network model and the algorithm learns and improves its accuracy over time.

Neural network data preprocessing

Neural networks require the input variables to be in numeric form. In the HMO dataset, most of our variables, namely ‘location’, ‘location\_type’, ‘Exercise’, Smoker’, ‘yearly\_physical’, ‘Hypertension’, ‘Gender’, ‘Education\_level’, ‘Married’, are categorical. Since Neural networks only consider pre processed data, unlike linear and logistic regression models whose algorithm automatically converts categorical variables into dummy variables, we manually converted the categorical variables into a numeric format before using it as input for the model. Finally, we used the following variables as an input for the model – ‘Age’,’ bmi’, ‘children’, ‘smoker’, ‘exercise’ and ‘hypertension’. We created dummy variables for categorical variables and then left out one class of each dummy variable to prevent redundant information.

Neural network with two nodes in one layer

#### **A close-up of a number Description automatically generated**

Neural network with three nodes in one layer.

A close-up of a number

Description automatically generated

With three nodes one layer neural network models, we made 2069 correct predictions, which consisted of 1806 correct non-expensive prediction and 263 correct expensive prediction, out of total 2275 testing observations. The predictive accuracy is 0.9095 and with 0.8969 to 0.9209 accuracy with 95% confidence interval. From the testing result, we have the basic conclusion that the three nodes model performed slightly better than the two node model.

Neural network with four nodes in one layer.

A screenshot of a computer

Description automatically generated

Association Rules

To generate the association rules for a particular dataset, important prerequisites are that the variables have to be of data type factor and the dataset needs to be converted to transaction. Meanwhile, unlike other predictive models, association rule analysis focuses more about data patterns appearing together. So we are not going to split the data into training and testing. Instead we will put all the data into the association rule model and find the rules with most instances. After all necessary data preprocessing for association rule analysis, we generated rules with a Support value of 0.05 and Confidence value of 0.70, we obtained 19 association rules that help us understand what the important variables are when predicting expensive clients.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

For the first rule, we have our determining factors and their subsequent values as given: expensive = 1, smoker = yes + exercise=Not-Active. This rule is describing the expensive clients who are smokers and are not active with exercise. This type of clients had a support of approximately 11% (0.109), which means that they were in the dataset for about 11% of the time, and had a confidence of 0.761, This rule has the highest Lift value as 3.801, which also supports an association between our antecedent and consequent. Smokers with not-active exercise status are the most possible clients that will make expensive cost. For the second rule, we had identical factors as the first rule with the addition of a new factor married = married. This had a Support of approximately 6% (0.056) as well and confidence of 0.707 and a slightly lower Lift value of 3.533, which still indicating an association between the smoker, not-active exerciser and expensive client, but this time, we also know those clients are possibly married. In this way, we could observe the rules and found a lot of common factors in most of them, such as the smoker, exercise as Not-Active, hypertension as 0, and yearly\_physical as No, which means a lot of expensive clients share the same life patterns, including smoking, not active exercise status, not having hypertension and not going to yearly physical check. Lowering the confidence level value and generating more rules could give us a deeper idea of more factors that led to expensive cases.

**Decision Tree**

Decision Trees are versatile Machine Learning algorithms which can perform both classification and regression tasks. We can visualize each decision made which makes decision tree a great tool for decision analysis. To train a decision tree we are splitting the data into train and test set. Since its easy to understand decision visualization for classification models we are classifying the cost into expensive as 1 and non-expensive as 0. We get the following decision tree plot for our model:

**A diagram of a number of people

Description automatically generated**

The above model gives the decision based on the most correlated variables age,BMI,exercise,smoker.It is clear that out of the total population the probability of a person being non-expensive is 80 percent while the remaining 20% are expensive if the person is a smoker and their probability is 67 percent. And as the node branches to smokers, it checks the BMI of the person if the BMI is less than 30 and smoker then the probability of that person's health cost being non-expensive is 39 percent and they cover 9 percent of the population under smokers. While if the person has a BMI above 30 the possibility that the person's medical cost will be expensive is 91 percent and 11 percent out of the total expensive population . And further observations from under the node BMI less than 30 suggest that a person being less than 37 years old will be inexpensive and the probability of that occurrence is 13 percent and they cover 5 % of the total expensive pool. While a person above 37 years with BMI less than 30 might be expensive and the probability of that occurrence is 63% and they include 5% of the population. Further analysis from the tree suggests that people with smoking issues, BMI less than 30, and who are in their late thirties or above and don't have an exercise regimen might be expensive and the possibility of that is 79 percent and they cover 4% out of the non-exercise population.

**A screenshot of a computer

Description automatically generated**

With the decision tree we made an accuracy of 91.12% and No Information Rate of 0.8105. For 95% confidence intervals the predictive accuracy ranges between 89.88 % to 92.26% .

Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. The Random forest for the dataset generated 500 trees. WIth 3 variables tried at each split.

**A black and white text on a white background

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**SVM(Support Vector Machine)**

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression. For building the support vector machine we divided the dataset into train and test taking 70 percent in train set and remaining 30 percent in test set. The model is built to predict whether a person’s medical cost is expensive or inexpensive, by the expensive variable as our dependent variable. We obtained the following finding:

**Iteration 1:**

For iteration 1 we are considering all variables and it gives as an accuracy of 90.4 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8913 to 0.916

**A screenshot of a computer

Description automatically generated**

**Iteration 2:**

For this iteration we are considering variables which have correlation such as age,BMI,children,Smoker, Yearly Physical, Exercise,Hypertension and Marriage Status. The SVM for these variables gave an accuracy of 90.73 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8946 to 0.9189.

A screenshot of a computer

Description automatically generated

**Iteration 3:**

For iteration 3 we removed marriage status variable and got an accuracy of 90.68 percent with NIR of 0.8105 and 95 percent confidence interval in the range of 0.8941 to 0.9184

A screenshot of a computer

Description automatically generated

**MODEL COMPARISON**

**A table with text and numbers

Description automatically generated**

### Conclusion:

With goals of understanding expensive clients and reducing the cost, we found that the clients who tend to be top 20% in cost have the characteristics of the following 1. Age has a positive impact on the cost, which means the older the clients are, the more possible for them to be an expensive client 2. Higher bmi also leads to higher costs. Body mass index (BMI) is a person’s weight in kilograms divided by the square of height in meters. BMI is an inexpensive and easy screening method for weight categories—underweight, healthy weight, overweight, and obesity. The bmi value bigger than 25.0 usually indicates overweight. 3. For other health indicators, clients who smoke, have hypertension, not being active with exercise, not going to regular physical checks usually have higher medical expenditures.

### Learning Outcomes:

* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
* Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
* Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

#### FudgeInc – Datawarehouse Implementation

**Course:** IST 722 Data Warehouse

**Project Description:**

To create a Datawarehouse that merges two business entities one is Fudgemart which is a retail business enterprise, and the other is FudgeFlix which is a rental DVD store. With the merged Datawarehouse we have to create a MOLAP (Multidimensional online Analytical Processing) cube which serves as the input for PowerBI dashboard.

**Data Warehouse Architecture:**

**A diagram of a diagram of a structure

Description automatically generated**

### Star Schema

A screenshot of a computer

Description automatically generated

Fact table: FactSales Dimension tables: DimProduct, DimDate, DimCustomer. Advantages: Performance: This star schema improves performance Simplicity: Easy to understand and write queries for

### OLAP Cube

A screenshot of a computer

Description automatically generated

Why a OLAP Cube?

* Serves a Multidimensional platform which helps to combine data into organized structures making it easy for analysis
* Fast Performance in terms of Query helping in aggregating data quickly. navigate intuitively.
* Hierarchical Relationships making it easy for users to

### Visual Analysis:

A screenshot of a computer

Description automatically generated

A screenshot of a map

Description automatically generated

### Recommendations

Based on OLTP

The OLTP Sources - especially FudgeMart must improve their customer city conventions. We can notice that city “Cleves” is mapped in Germany. There is no indication whether customers are worldwide or exclusive to the USA.

Based on Visual Analytics

* We could see the overall sales revenue shows a downward trend over the years. FudgeInc has to optimize the pricing strategy. Strategies like focus on high demand products ,discontinuing underperforming ones and understanding the market trends might help.
* We observe a downward trend on sales during Q4. Solution is to try marketing discounts, promotions and advertising strategies. Like Holiday themed products/bundles, Early bird promotions etc

### Learning Outcomes achieved from this course and projects :

* Collect, store, and access data by identifying and leveraging applicable technologies using the datawarehouse designing process
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle- this is achieved by the PowerBI dashboard visualization which gave recommendations to the stakeholders.

#### Los Angeles Crime Analysis using Tableau

**Course:** IST 737 Visual Analytics Dashboard

**Project Description:**

A tableau dashboard that focuses on crime in Los Angeles city, the projects gives a clear insight on victims affected, type of crime and area of crime. This dashboard will be helpful resource for entities like the Department of police, City government and also for the people who are in LA or planning to move to LA.

Dataset used: <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8>

**Dashboard 1:**

**A close-up of a graph

Description automatically generated**

This dashboard presents a crime analysis across various dimensions. On the left, there is a list of frequent crimes, with "Vehicle - Stolen" being the most common, occurring 44,445 times. This section also includes crimes like battery, vandalism, burglary, and theft with their respective counts. Below, there's a "Crime Count by Day of the Week" heatmap, indicating that crime is most frequent on Fridays and least on Mondays. On the right, a "Crime Count by Race" bubble chart shows racial distribution of crime counts, with Hispanic and Black communities having higher counts compared to Asian and other races. Below, the "Crimes by Victims" bar graph displays the distribution of different crime types such as assault, battery, burglary, and robbery, showing the counts of victims by crime type with a color-coded key. This visual representation allows for quick insights into patterns and distributions of crime data.

**Dashboard 2:**

**A screenshot of a computer screen

Description automatically generated**

This dashboard provides a location-based crime report with interactive elements for detailed analysis. Users can click on different locations listed on the left to view specific data. The report includes a bar chart showing the premises of crimes with "Street" being the most common location, followed by single-family dwellings, and others. The "Common Weapon Used" chart indicates that "Strong-arm" (hands, fists, feet, etc.) is the most common, with handguns and semi-automatic pistols also noted. Below, a heatmap illustrates crime distribution across the Southwest location, with numerous incidents highlighted. Lastly, the "Affected Gender" chart breaks down victimization by age and sex, showing a higher incidence in younger male individuals, with a peak in the 20-24 age range for both males and females.

**Dashboard 3:A screenshot of a graph

Description automatically generated**

The dashboard displays a "Crime Time Report" with both yearly and monthly crime data, as well as crime frequency by hour of the day. The "Yearly Crime Report" line graph shows an increase in crime from 199,409 incidents in 2020 to a peak of 234,227 in 2022, followed by a significant decrease to 177,610 in 2023, which is because the complete data of 2023 is not available as of November. The "Monthly Crime Report" line graph beneath reveals seasonal trends with crime peaking in July and generally declining towards the end of the year. The bar graph on the right, "Crime during hour of the day," indicates that crime occurrences peak sharply at midnight and have a general uptrend in the evening hours, tapering off in the early morning hours.

**Dashboard 4:**

**A screenshot of a graph

Description automatically generated**

This dashboard provides a comprehensive overview of crime investigation status, ongoing investigation percentages, and arrest comparisons. The left section, "Investigation Status for Crime Types," shows a bar graph with various crimes and their corresponding clearance rates, indicating how many cases have been solved or are progressing, with "Vehicle - Stolen" having the highest clearance rate. The middle section, "Investigation Report," features a donut chart revealing that 80.05% of cases have been under investigation since 2020. The right section, "Adult Arrest and Juvenile Arrest Comparison based on Type of Weapon Used," presents a bar graph contrasting the number of adult and juvenile arrests for different crimes. Adult arrests for assault with a deadly weapon have the highest count, significantly more than juvenile arrests, with other crimes following in a descending order.

### Key learning outcomes achieved from this course and project:

* Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads.
* Apply visualization and predictive models to help generate actionable insight.

#### Lacrosse team data pipeline

**Course:** IST 769 Advanced Big Data Management (Ongoing coursework)

Since this is an ongoing course but has projects that can help me in my career learning curve. I would like to showcase this on my first draft and will update it on my final draft with my final project. The course work requires us to do 3 projects that focuses on Big data pipeline tools that are widely used in the Data Engineering Industry. I have currently completed my 1st project which is my mid term so I am attaching that in my first draft.

**Project Description:**

This datapipeline focuses on processing a Lacrosse game’s live data, that creates box score and updates the database tables when game is over.

### Databases**:**

• **mssql** : Microsoft SQL Server Database that stores player and teams reference data. The database is called ‘sidearm’ and the tables are ‘players’ and ‘teams’.

• **minio**: An S3 compatible object store that contains the live gamestream data. It is stored in minio/gamestreams bucket.

• **mongodb** – A mongodb database that stores the game streams real-time box store so the web developers can create a page from the data. The box score is written to mongo/sidearm/boxscores collection.

## Analytical Tools:

• **drill** - An instance of Apache Drill that can be used to query the databases. The drill-storage-plugins folder contains the configuration files for the databases. You will need to modify these with specifics for them to work.

• **jupyter** - An instance of Jupyter Lab that can be used to write PySpark code. The work folder contains the Start.ipynb that demonstrates the base spark configuration.

## The Problem:

The objective is to create a data pipeline which processes a simplified version of an in-game stream from a simulated a lacrosse game. The game stream has been simplified to only process goals scored. There are two parts to this problem:

1. At any point while the game is in progress, the game stream should be converted into a JSON format so the web developers can use it to create a box score page on a website. This JSON should be written to the mongodb/sidearm/boxscores collection, and should contain all the data necessary to display the box score page
2. When the game is over, the player and team reference data should be updated to reflect the team records and player statistics after the competition has ended. Normally you would update the mssql tables, but for this exam you will create new tables with the updated data, players2 and teams2 respectively. This is mostly because spark does not support row-level updates. In a real world scenario, you would write an SQL script on mssql to update the tables. from the changes in the players2 and teams2 tables, but that is outside the scope of this exam.

**Data Pipeline Constructed:A diagram of a program

Description automatically generated**

### Key Learning Outcomes Achieved from this course and project:

* Collect, store, and access data by identifying and leveraging applicable technologies
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
* Use programming languages such as R and Python to support the generation of actionable insight

### Conclusion:

Throughout my two-year journey pursuing a Master's degree in Data Science, I've come to realize that the field extends far beyond mere data manipulation and analysis. It encompasses the critical ability to make unbiased decisions, select and integrate appropriate tools to maintain data performance, and uphold the responsibility of ensuring accessibility to insights for all who seek them.

My time at Syracuse University and the iSchool has been transformative, providing me with a rich and fruitful learning experience. I am grateful for the invaluable knowledge and skills acquired during this period, which have not only deepened my understanding of Data Science but also equipped me with the tools to navigate its complexities effectively.