Portfolio Milestone

Nadia Paquin
June 16, 2024

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied Data Science
Syracuse University, School of Information

Portfolio Roadmap

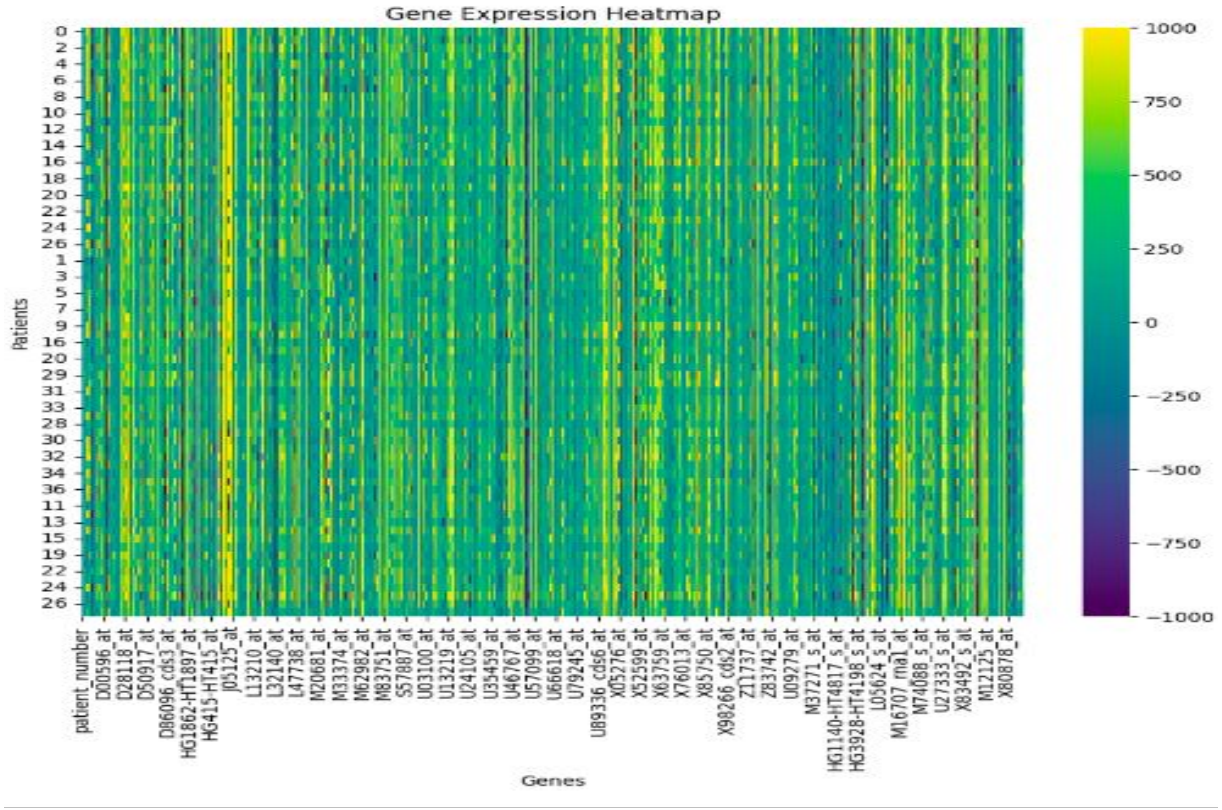
- Portfolio components
  - Four projects demonstrating proficiency in MSADS learning outcomes
  - Written report
  - Presentation
  - Github repository
- Learning outcomes of the MSADS program
- Project descriptions
- Conclusions, references, contact information

MSADS Learning Outcomes

- Collect, store, and access data by identifying and leveraging applicable technologies.
- Create actionable insight across a range of contexts 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.
- Apply ethics in the development, use and evaluation of data and predictive models.

IST718 Big Data Analytics (Python)

- **Project:** Differentiate between acute myeloid leukemia (or AML) and acute lymphoblastic leukemia (or ALL) using gene expression data
- **Data:** 7131 genes x 72 patients (500,000+ data points)
- **Method:** Clean, visualize, model



ALL vs. AML Classification

Model	Accuracy
K-means	0.853
SVM	0.941
Logistic Regression	1.00
Naive Bayes	0.912
Random Forest	0.912

IST664 Natural Language Processing (Python)

- **Project:** Classify spam emails using NLTK
- **Data:** 1500 spam emails and 3672 non-spam “ham” emails
- **Method:** Train NLTK’s naive bayes classifier and test
  - Variations: vocabulary size, corpus size, stop and negation words, POS tagging, n-grams, ratio of words to punctuation.

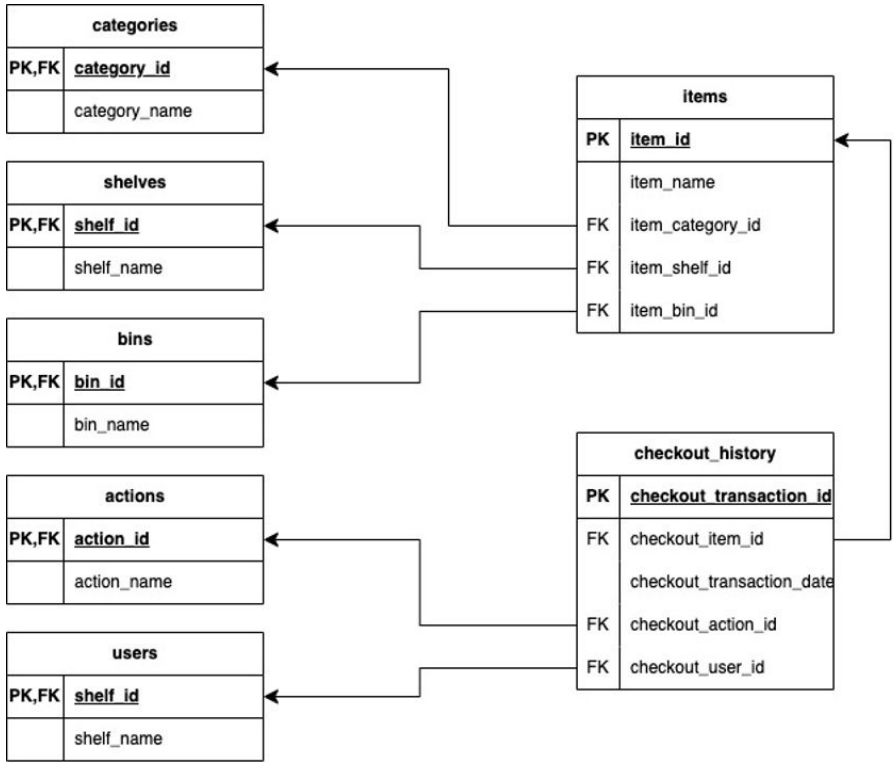
Model	Accuracy
Top 2000 most common words	0.954
Top 1000 most common words	0.953
Top 500 most common words	0.956
Top 100 most common words	0.954
250 input files	0.942
100 input files	0.880
25 input files	0.720
5 input files	0.400
Removing stop words and negation	0.950
Top 500 bigrams	0.950
Word to punctuation ratio	0.592
POS tagging	0.953
Removing stop words + Top 500 bigrams	0.952
Top 2000 word frequency	0.945

IST707 Applied Machine Learning (R)

- **Project:** Build a model to diagnose patients from symptoms
- **Data:** 132 symptoms x 4290 patients
- **Method:** Decision trees, Naive Bayes, Random Forest and KNN models

Model	Accuracy
Decision tree	0.74
Naive Bayes	0.89
Random Forest	0.93
KNN	0.99

IST659 Database Administration



Conclusions

- 4 Projects:
  - 2 with focus on standard data cleaning, explore, visualize, and model pipeline in both R and Python with large datasets and a variety of machine learning models to produce actionable insights
  - 1 with focus on parsing and classifying text data
  - 1 with focus on database building and management systems in SQL

References

Golub, T R et al. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.” Science (New York, N.Y.) vol. 286,5439 (1999): 531-7.

KAUSHIL268 (2020). Disease Prediction Using Machine Learning [Data set]. Kaggle. https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning/data

V. Metsis, I. Androutsopoulos and G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?". Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006.

Github link: https://github.com/nadiapaquin/MSADS_Portfolio/tree/main