About the Project

**The Problem Statement**

The AI model attempts to answer the following questions in an attempt to quell the overburdened healthcare system Canada is currently experiencing:

* Can a professional’s expertise be reflected in a single model, allowing for higher volume of diagnoses?
* What is an inaccuracy in this model, what is the risk, and how can it be accounted for?
* Can this model and data be used proactively for resource allocation & funding decisions?

**The Model**

This project uses diagnoses data to create clear links between symptoms and diseases, as well as provides recommendations on how to diet and exercise according to the predicted disease. It also supports photoimaging of the lungs to flag if an infection is detected.

The data is first loaded in to Python with the Pandas library and cleaned to ensure optimal performance of the database and model; outputs are validated using the Matplotlib and Seaborn libraries.

The data is then loaded in and stored in a database; the AI model pulls from this database, and uses a Single Vector Machine model to equate symptoms with diseases. It does this via a flask-hosted app where the user can type in up to four symptoms – the app has input validation and synonym dictionaries to help guide the user.

**Data Sources**

**Kaggle Diagnoses Data**

<https://www.kaggle.com/datasets/noorsaeed/medicine-recommendation-system-dataset/data>

**CBC Healthcare System is Overburdened**

<https://www.cbc.ca/news/canada/toronto/healthcare-staff-union-survey-1.7073334>

**Evaluation Standards**

The model will be evaluated on a points-system with 100 potential points.

**Data Model Implementation (25 points)**

* A Python script initializes, trains, and evaluates a model (10 points)
* The data is cleaned, normalized, and standardized prior to modeling (5 points)
* The model utilizes data retrieved from SQL or Spark (5 points)
* The model demonstrates meaningful predictive power at least 75% classification accuracy or 0.80 R-squared. (5 points)

**Data Model Optimization (25 points)**

* The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself (15 points)
* Overall model performance is printed or displayed at the end of the script (10 points)

**GitHub Documentation (25 points)**

* GitHub repository is free of unnecessary files and folders and has an appropriate .gitignore in use (10 points)
* The README is customized as a polished presentation of the content of the project (15 points)

**Group Presentation (25 points)**

* All group members speak during the presentation. (5 points)
* Content, transitions, and conclusions flow smoothly within any time restrictions. (5 points)
* The content is relevant to the project. (10 points)
* The presentation maintains audience interest. (5 points)

**Conclusion**

The model concludes that there is a very large opportunity for the use of AI/ML in the Healthcare industry; while risk of inaccuracies necessitate some level of human supervision, high initial accuracy rates over a large number of requests indicate that these tools can be used as a filtering mechanism to ‘bucket’ diagnoses, flagging responses with similar historically inaccurate results.