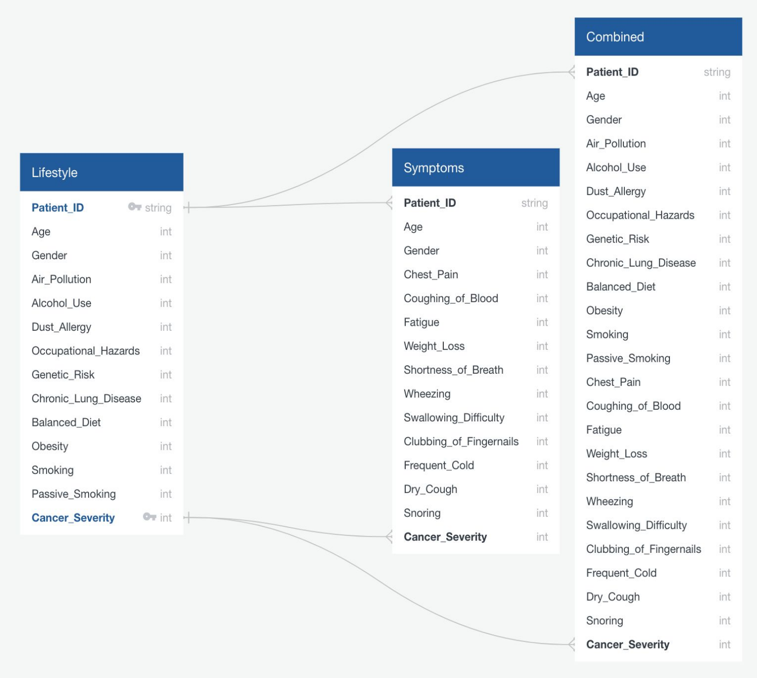
Group 6 - Cancer Predictions

Raima Ghosh - Shane Abbley - Janell Napper - Amit Choksi

Which lifestyle choices determine the severity of a cancer diagnosis?

Data and database

- We obtained our cancer data from Kaggle.
- Our data contains symptom severity and lifestyle choices ranked from 1 to 8, and cancer severity ranked from 0 - 3
- We created a SQLite database containing the following three tables:



Lifestyle	
Patient_ID	string
Age	int
Gender	int
Air_Pollution	int
Alcohol_Use	int
Dust_Allergy	int
Occupational_Hazards	int
Genetic_Risk	int
Chronic_Lung_Disease	int
Balanced_Diet	int
Obesity	int
Smoking	int
Passive_Smoking	int
Cancer_Severity	int

Symptoms	
Patient_ID	string
Age	int
Gender	int
Chest_Pain	int
Coughing_of_Blood	int
Fatigue	int
Weight_Loss	int
Shortness_of_Breath	int
Wheezing	int
Swallowing_Difficulty	int
Clubbing_of_Fingernails	int
Frequent_Cold	int
Dry_Cough	int
Snoring	int
Cancer_Severity	int

Combined	
Patient_ID	string
Age	int
Gender	int
Air_Pollution	int
Alcohol_Use	int
Dust_Allergy	int
Occupational_Hazards	int
Genetic_Risk	int
Chronic_Lung_Disease	int
Balanced_Diet	int
Obesity	int
Smoking	int
Passive_Smoking	int
Chest_Pain	int
Coughing_of_Blood	int
Fatigue	int
Weight_Loss	int
Shortness_of_Breath	int
Wheezing	int
Swallowing_Difficulty	int
Clubbing_of_Fingernails	int
Frequent_Cold	int
Dry_Cough	int
Snoring	int
Cancer_Severity	int

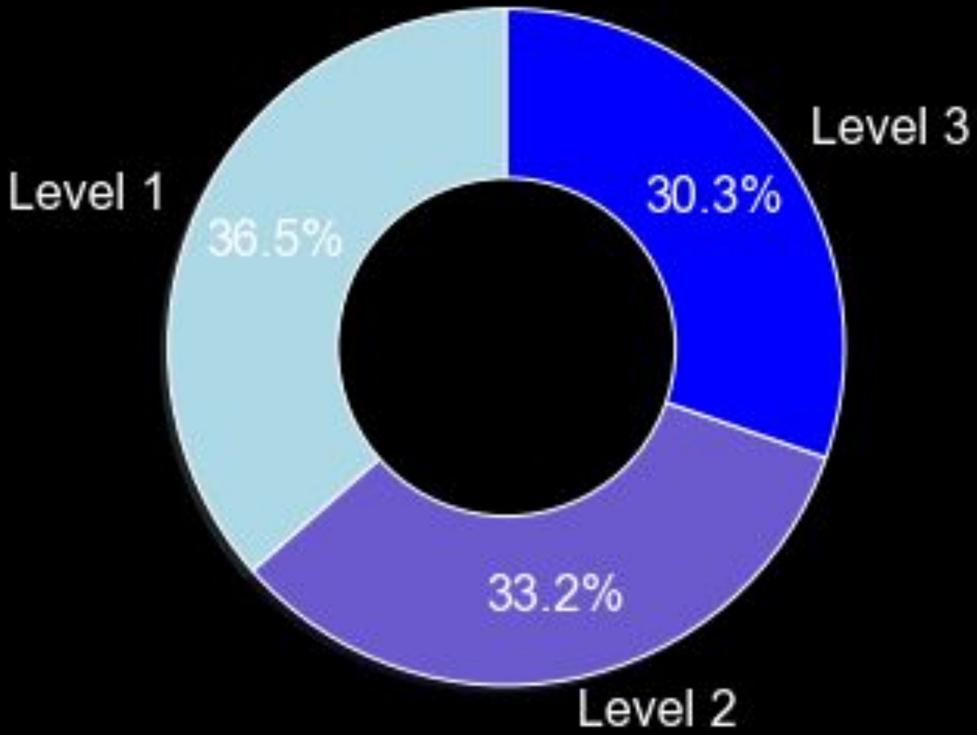
After looking over this data, we wanted to answer these three questions.

1. Which lifestyle choices are connected to a higher incidence of cancer?

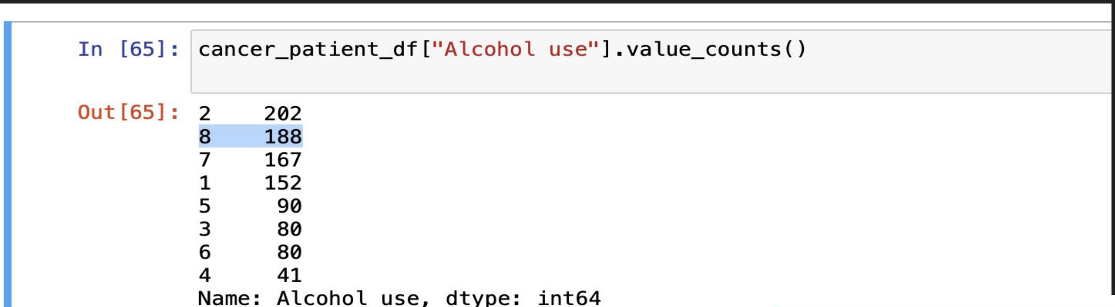
2. Are there multiple lifestyle factors that are associated with a higher incidence of cancer?

3. Is there any lifestyle choice associated with more severe outcomes of Cancer?

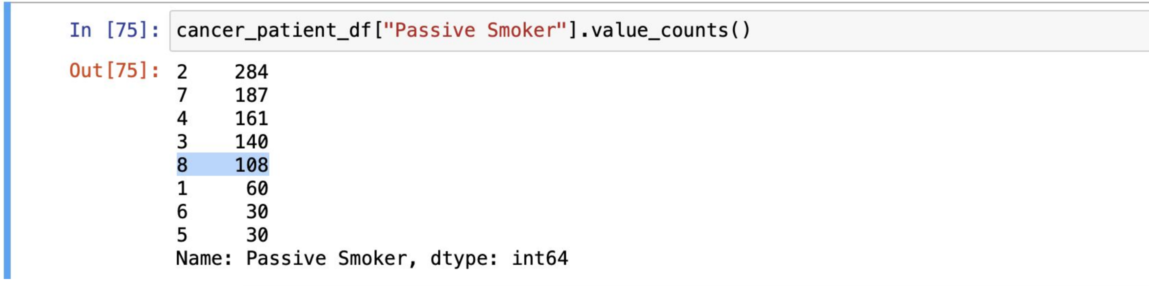
Cancer Severity by Level



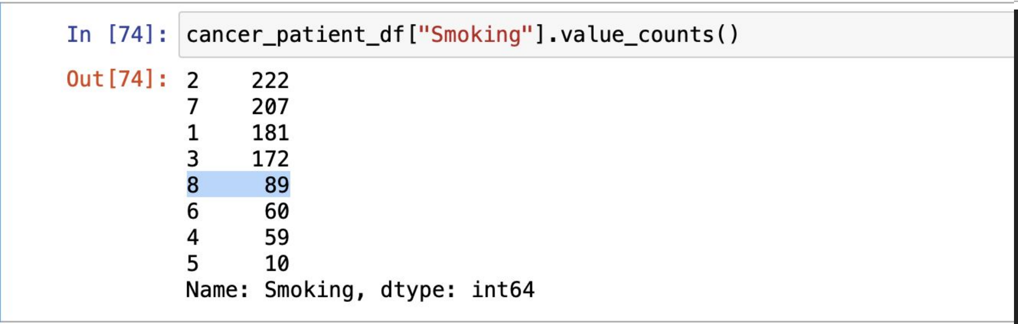
Top Three Columns with the Highest usage value (8)



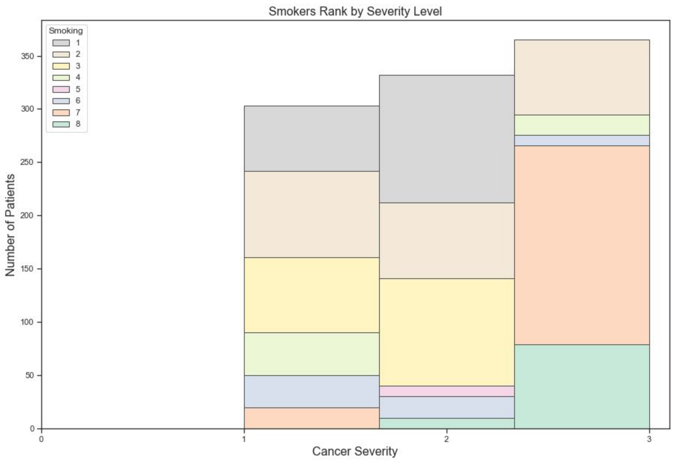
```
In [65]: cancer_patient_df["Alcohol use"].value_counts()

Out[65]: 2    202
         8    188
         7    167
         1    152
         5     90
         3     80
         6     80
         4     41
         Name: Alcohol use, dtype: int64
```



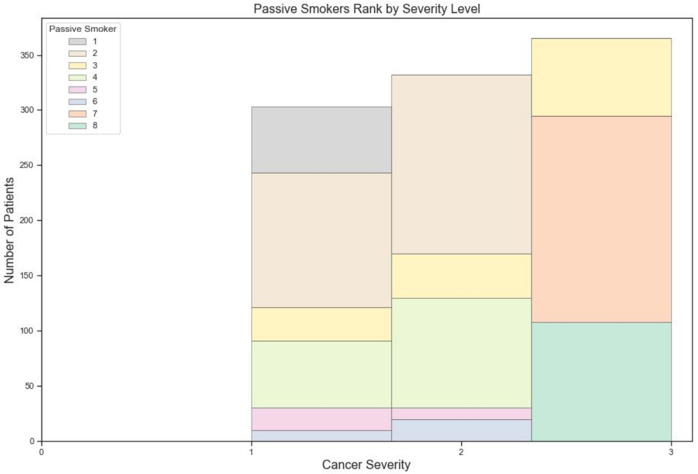
```
In [75]: cancer_patient_df["Passive Smoker"].value_counts()

Out[75]: 2    284
         7    187
         4    161
         3    140
         8    108
         1     60
         6     30
         5     30
         Name: Passive Smoker, dtype: int64
```



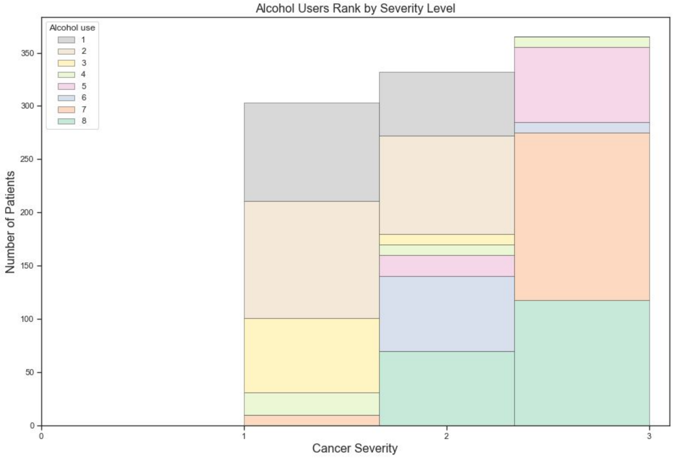
```
In [74]: cancer_patient_df["Smoking"].value_counts()

Out[74]: 2    222
         7    207
         1    181
         3    172
         8     89
         6     60
         4     59
         5     10
         Name: Smoking, dtype: int64
```



Smokers Rank by Severity Level



Passive Smokers Rank by Severity Level



Alcohol Users Rank by Severity Level

Model Overview

- Model of Choice : Logistic Regression & SVM for the predicting the severity of Cancel as Low, Medium, or High
- We can compare the various results from running these algorithms
- Ran the model with different solvers, and different values of iterations.
- Got the best results with newton-cg solver (86%)
- Logistic Regression : This algorithm is used for classification problems in machine learning.
- Support vector machine : This algorithm separates the data points using a line , this line is chosen such that it will be furthermost from the nearest data points in 2 categories

Logistic Regression

It is a classification model which is used to predict the odds in favour of a particular event. The odds ratio represents the positive event which we want to predict, for example, how likely a sample has breast cancer/ how likely is it for an individual to become diabetic in future. It used the sigmoid function to convert an input value between 0 and 1. It can further be extended to multiple logistic regressionThe basic idea of logistic regression is to adapt linear regression so that it estimates the probability a new entry falls in a class. The linear decision boundary is simply a consequence of the structure of the regression function and the use of a threshold in the function to classify. Logistic Regression tries to maximize the conditional likelihood of the training data, it is highly prone to outliers. Standardization (as co-linearity checks) is also fundamental to make sure a features' weights do not dominate over the others.

Source :
https://www.geeksforgeeks.org/differentiate-between-support-vector-machine-and-logistic-regression/

SVM

It is a very powerful classification algorithm to maximize the margin among class variables. This margin (support vector) represents the distance between the separating hyperplanes (decision boundary). The reason to have decision boundaries with large margin is to separate positive and negative hyperplanes with adjustable bias-variance proportion. The goal is to separate so that negative samples would fall under negative hyperplane and positive samples would fall under positive hyperplane. SVM is not as prone to outliers as it only cares about the points closest to the decision boundary. It changes its decision boundary depending on the placement of the new positive or negative events.

The decision boundary is much more important for Linear SVM's – the whole goal is to place a linear boundary in a smart way. There isn't a probabilistic interpretation of individual classifications, at least not in the original formulation.

Source : https://www.geeksforgeeks.org/differentiate-between-support-vector-machine-and-logistic-regression/

Comparison Pros/Cons of using various models

Benefits of Logistic regression

- solving classification problem

- not used to find the best margin instead it can have different decision boundaries with different weights that are near the optimal point

- works with already identified independent variable

Cons

- vulnerable to overfitting

SVM

- tries to find the best margin that separates the classes that reduces the risk of error on the data

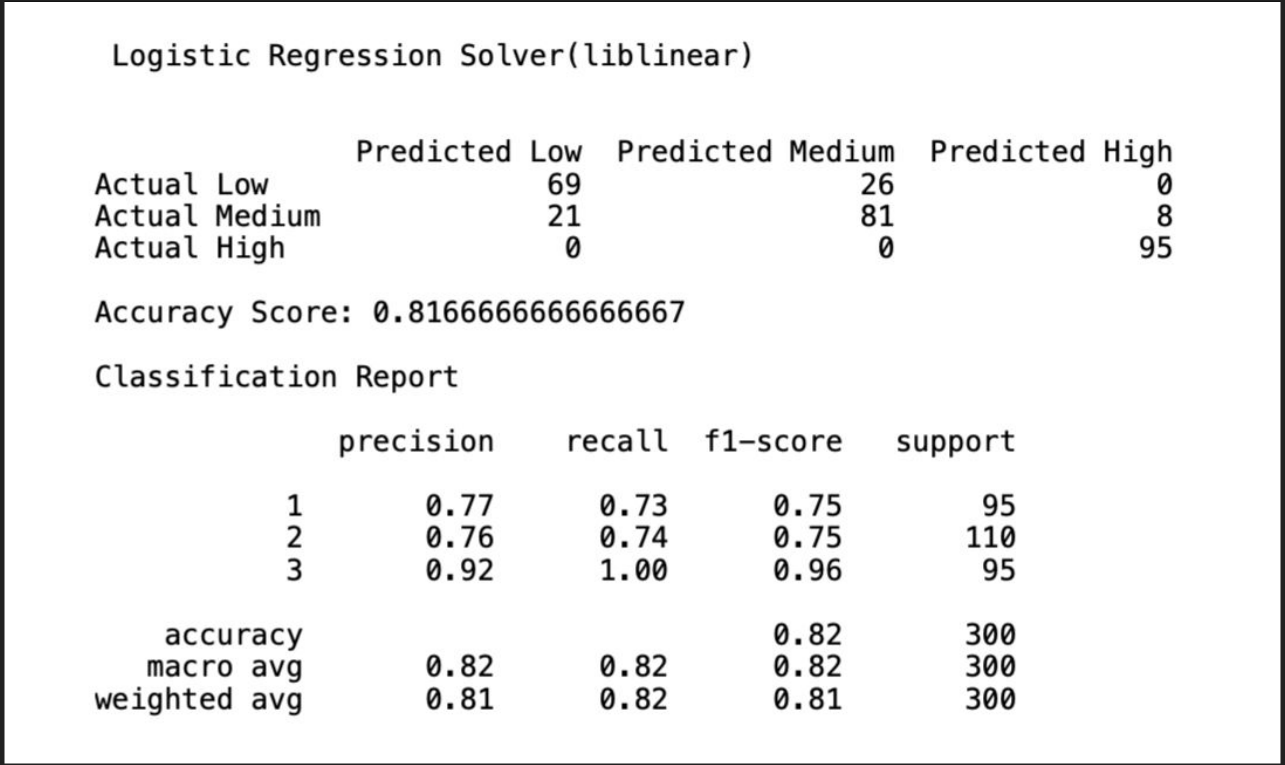
- risk of overfitting is less.

- gives the best prediction for the model under study.

Overfitting

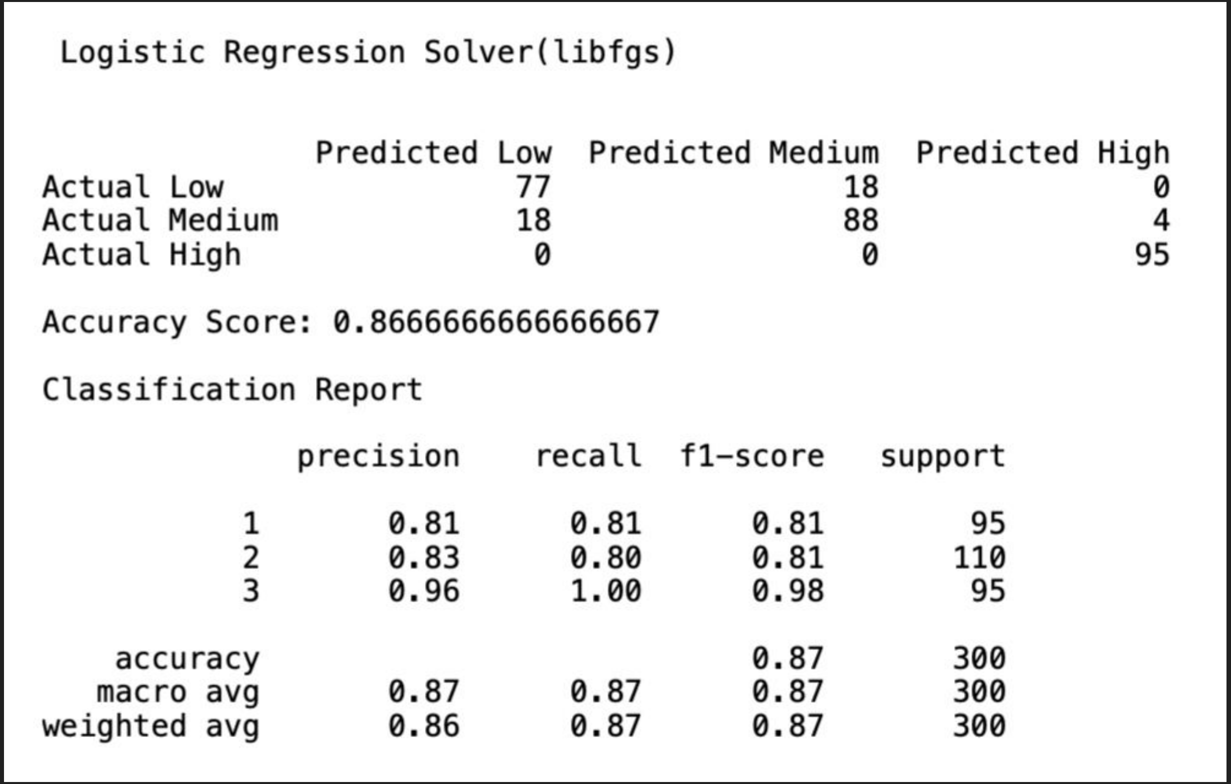
Overfitting is **an error that occurs in data modeling as a result of a particular function aligning too closely to a minimal set of data points**. This is not much applicable for the size of the data set we have but is one something of prime importance for huge unbalanced data sets.

Logistic Regression (liblinear)



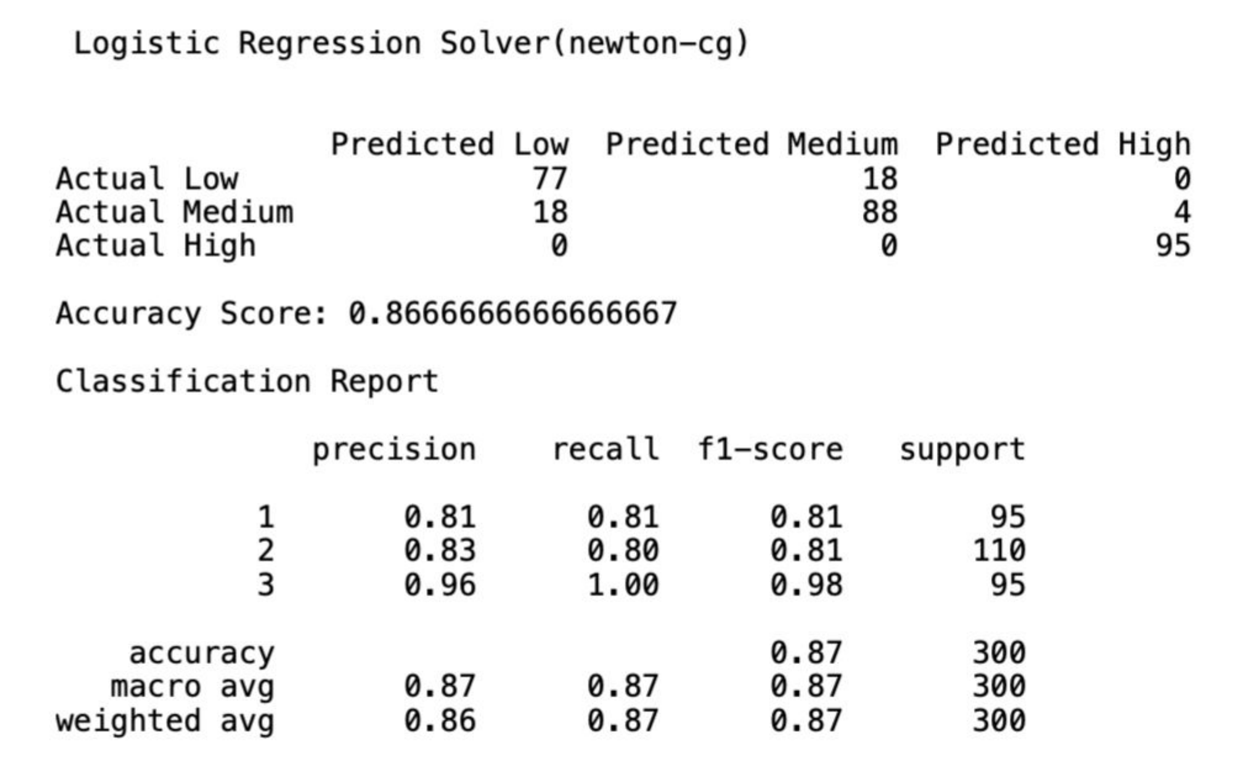
```
 Logistic Regression Solver(liblinear)


               Predicted Low  Predicted Medium  Predicted High
Actual Low                69                26               0
Actual Medium             21                81               8
Actual High                0                 0              95

Accuracy Score: 0.8166666666666667

Classification Report

              precision    recall  f1-score   support

           1       0.77      0.73      0.75        95
           2       0.76      0.74      0.75       110
           3       0.92      1.00      0.96        95

    accuracy                           0.82       300
   macro avg       0.82      0.82      0.82       300
weighted avg       0.81      0.82      0.81       300
```

Logistic Regression libfgs



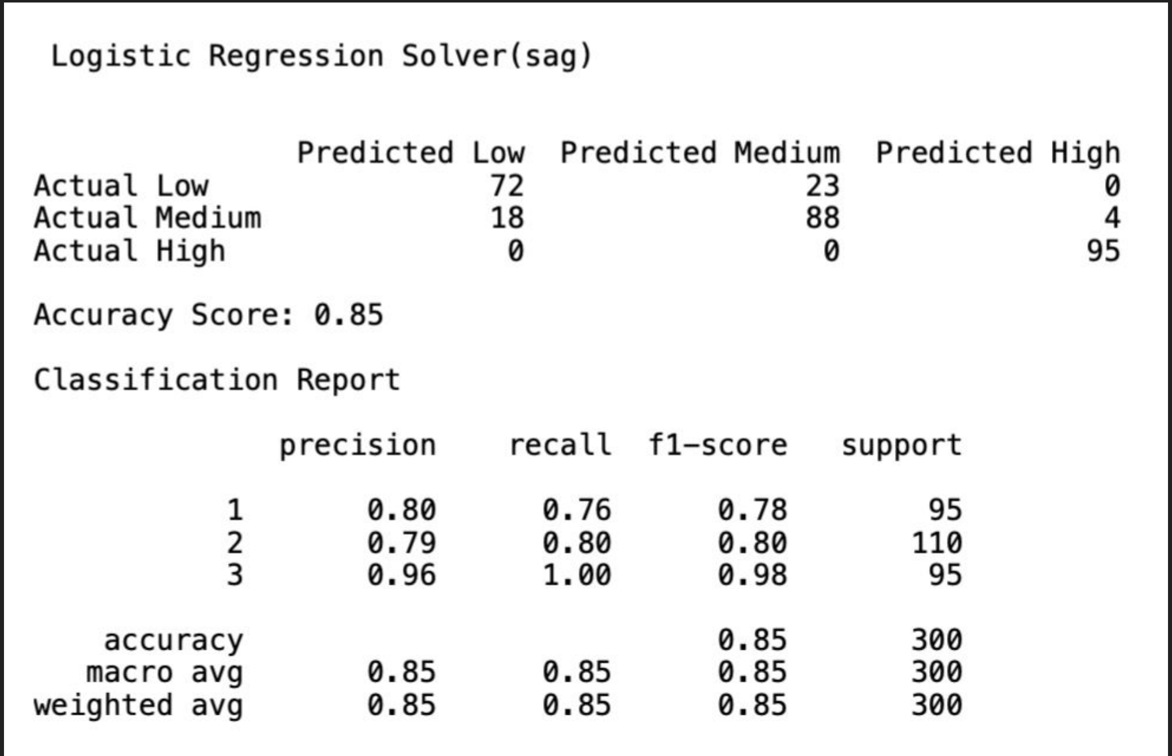
```
 Logistic Regression Solver(libfgs)


               Predicted Low  Predicted Medium  Predicted High
Actual Low                77                18               0
Actual Medium             18                88               4
Actual High                0                 0              95

Accuracy Score: 0.8666666666666667

Classification Report

              precision    recall  f1-score   support

           1       0.81      0.81      0.81        95
           2       0.83      0.80      0.81       110
           3       0.96      1.00      0.98        95

    accuracy                           0.87       300
   macro avg       0.87      0.87      0.87       300
weighted avg       0.86      0.87      0.87       300
```

Logistic Regression newton-cg



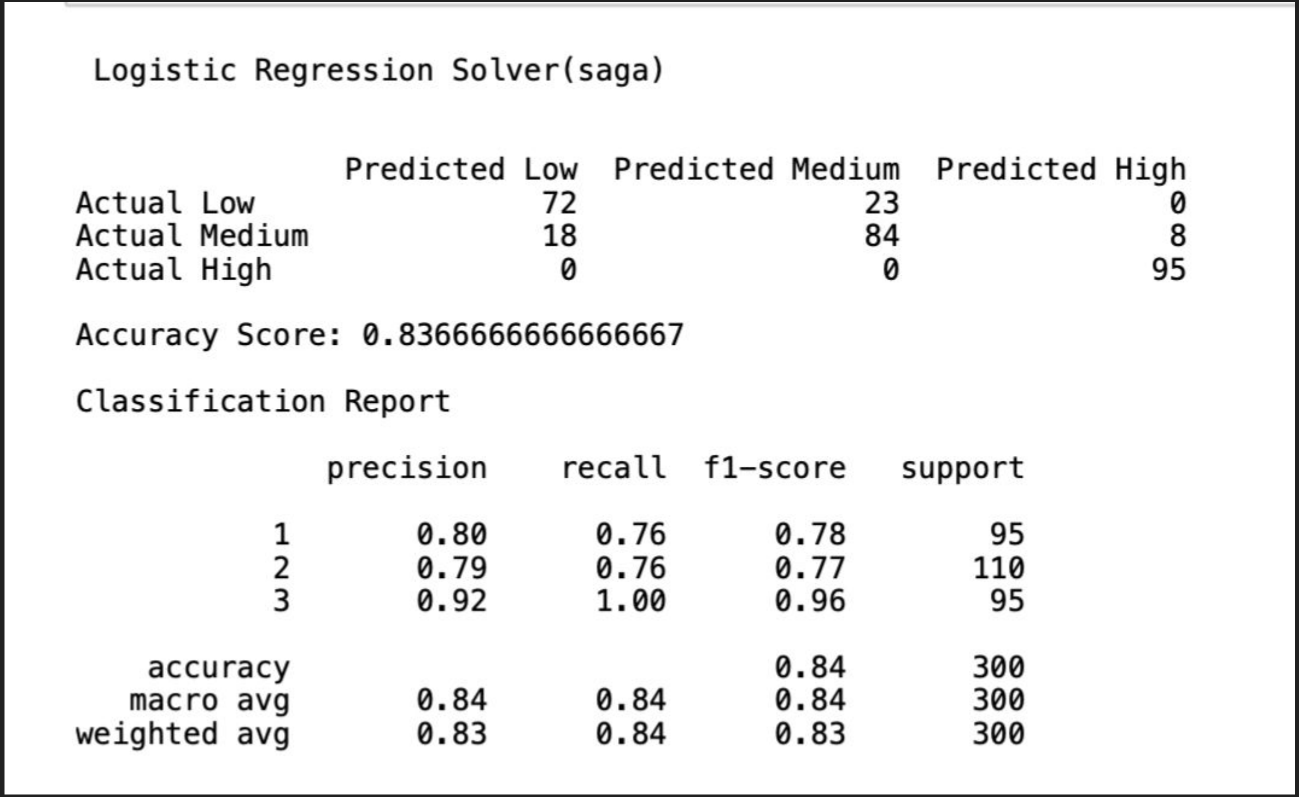
```
 Logistic Regression Solver(newton-cg)


                Predicted Low  Predicted Medium  Predicted High
Actual Low                 77                18               0
Actual Medium              18                88               4
Actual High                 0                 0              95

Accuracy Score: 0.8666666666666667

Classification Report

              precision    recall  f1-score   support

           1       0.81      0.81      0.81        95
           2       0.83      0.80      0.81       110
           3       0.96      1.00      0.98        95

    accuracy                           0.87       300
   macro avg       0.87      0.87      0.87       300
weighted avg       0.86      0.87      0.87       300
```

Logistic Regression sag



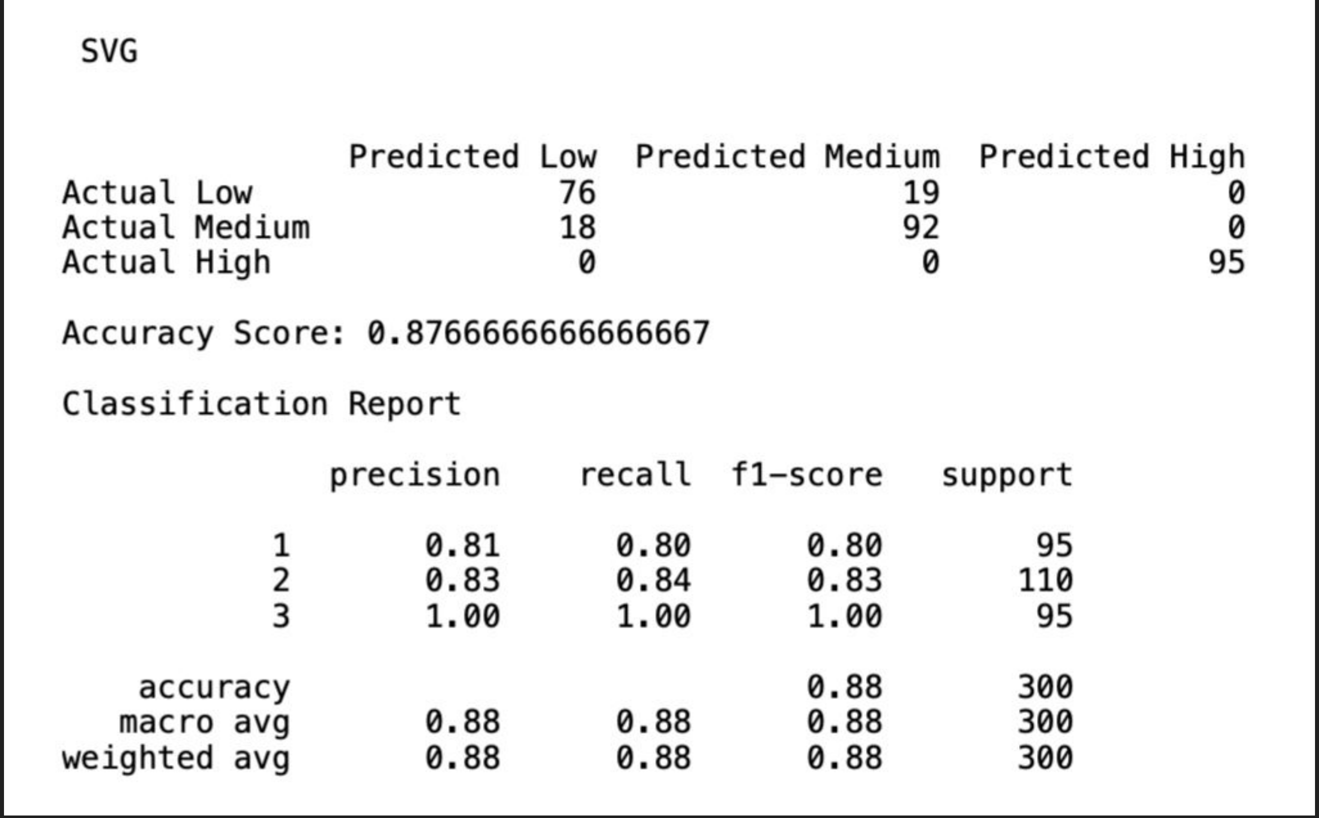
```
 Logistic Regression Solver(sag)


               Predicted Low  Predicted Medium  Predicted High
Actual Low                72                23               0
Actual Medium             18                88               4
Actual High                0                 0              95

Accuracy Score: 0.85

Classification Report

              precision    recall  f1-score   support

           1       0.80      0.76      0.78        95
           2       0.79      0.80      0.80       110
           3       0.96      1.00      0.98        95

    accuracy                           0.85       300
   macro avg       0.85      0.85      0.85       300
weighted avg       0.85      0.85      0.85       300
```

Logistic Regression saga



```
 Logistic Regression Solver(saga)


               Predicted Low  Predicted Medium  Predicted High
Actual Low                72                23               0
Actual Medium             18                84               8
Actual High                0                 0              95

Accuracy Score: 0.8366666666666667

Classification Report

              precision    recall  f1-score   support

           1       0.80      0.76      0.78        95
           2       0.79      0.76      0.77       110
           3       0.92      1.00      0.96        95

    accuracy                           0.84       300
   macro avg       0.84      0.84      0.84       300
weighted avg       0.83      0.84      0.83       300
```

SVG Algorithm (Best prediction)



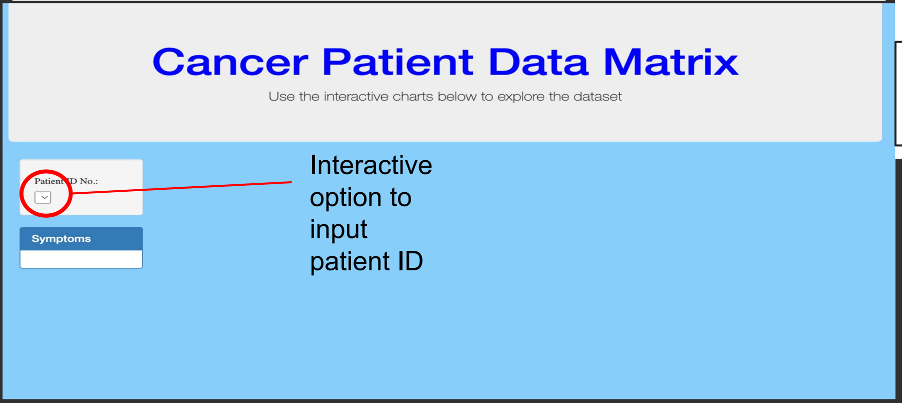
```
 SVG


              Predicted Low  Predicted Medium  Predicted High
Actual Low               76                19               0
Actual Medium            18                92               0
Actual High               0                 0              95

Accuracy Score: 0.8766666666666667

Classification Report

             precision    recall  f1-score   support

           1       0.81      0.80      0.80        95
           2       0.83      0.84      0.83       110
           3       1.00      1.00      1.00        95

    accuracy                           0.88       300
   macro avg       0.88      0.88      0.88       300
weighted avg       0.88      0.88      0.88       300
```

Dashboard

Raima Ghosh

Current View of the Dashboard

- Coded using d3.json with Bootstrap components
- Csv files will be cleaned using pandas and converted to json
- Will also display features that come out of ML model and preliminary data graphs
- Interactive input connected to Symptom info and bar graph



Cancer Patient Data Matrix

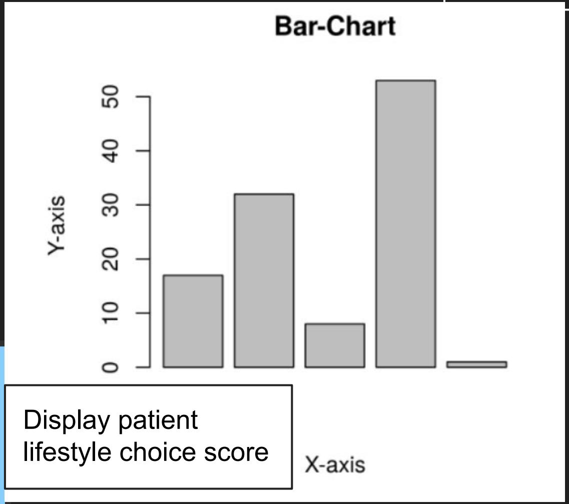
Use the interactive charts below to explore the dataset

Patient ID No.:

Interactive option to input patient ID

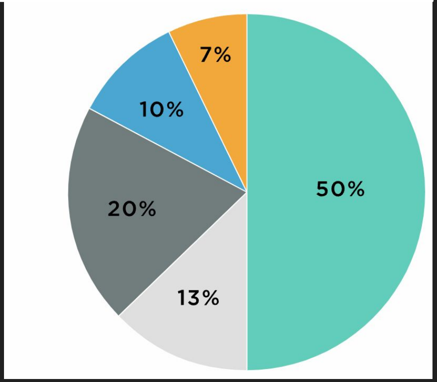
Another dashboard will allow user input of symptoms or lifestyle choices that will run a ML model to predict level of cancer risk. This will be deployed using Flask

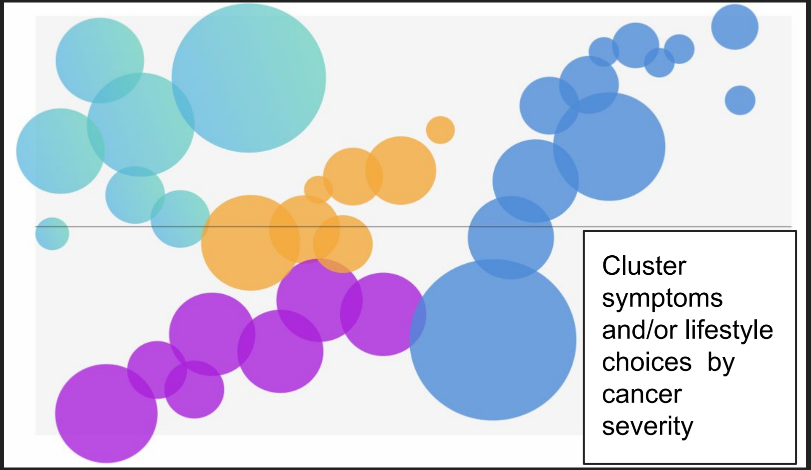
Potential Future Visualizations



Display patient lifestyle choice score

Classify lifestyle scores at different cancer severity





Cluster symptoms and/or lifestyle choices by cancer severity