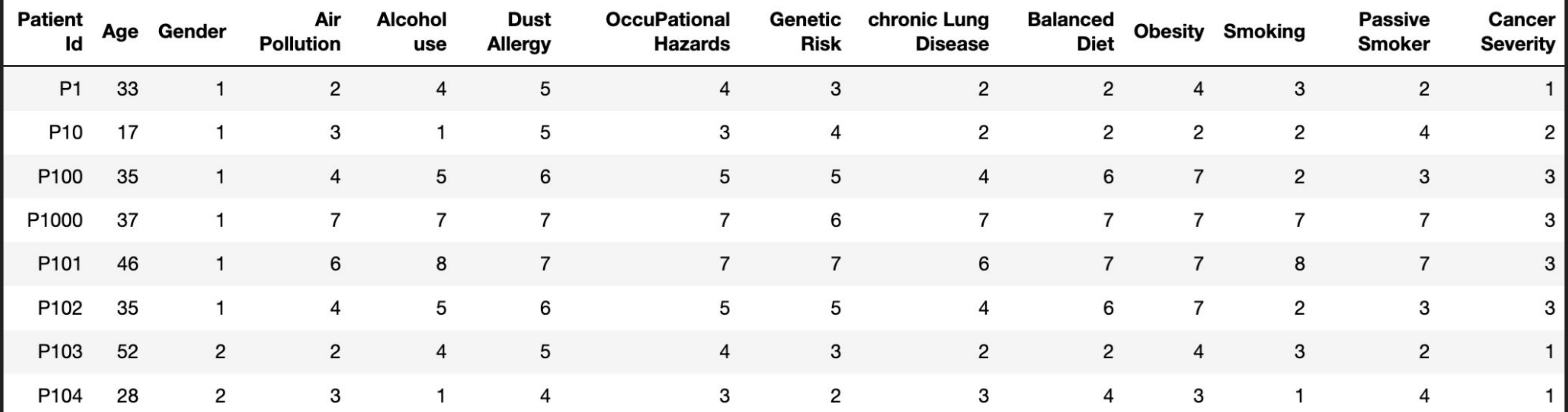
Group 6 - Cancer Predictions

Raima Ghosh - Shane Abbley - Janell Napper - Amit Choksi

Which lifestyle choices determine the severity of a cancer diagnosis?

Data and Data Cleaning

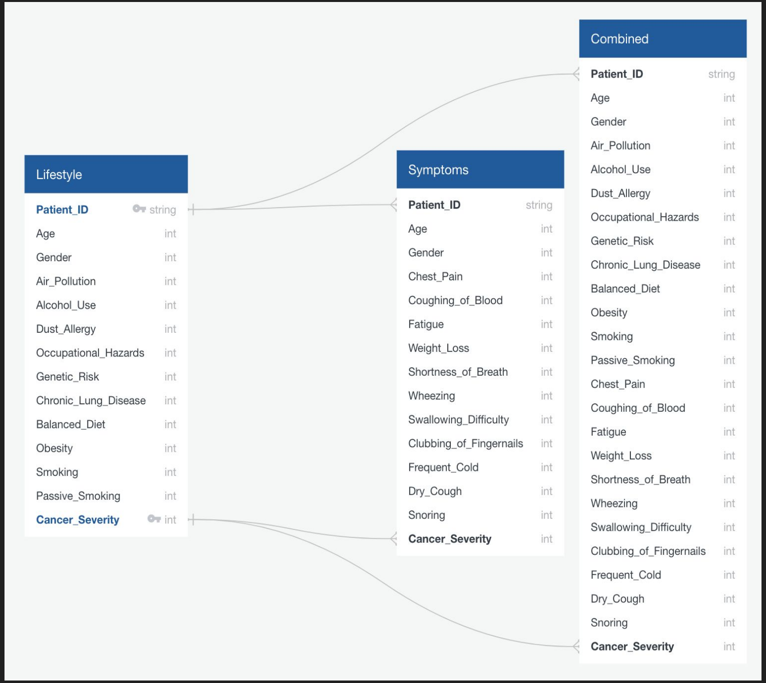
- We obtained our lung cancer data from Kaggle.
- Our data contains symptom severity and lifestyle choices ranked from 1 to 8 and cancer severity ranked as either high, moderate, or low.
- We cleaned the data using pandas in a jupyter notebook.
  - We checked for empty data.
  - We dropped all of the symptom columns and converted them to a separate data frame for our visualizations.
  - We converted the cancer severity to a numeric value.



Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Cancer Severity
P1	33	1	2	4	5	4	3	2	2	4	3	2	1
P10	17	1	3	1	5	3	4	2	2	2	2	4	2
P100	35	1	4	5	6	5	5	4	6	7	2	3	3
P1000	37	1	7	7	7	7	6	7	7	7	7	7	3
P101	46	1	6	8	7	7	7	6	7	7	8	7	3
P102	35	1	4	5	6	5	5	4	6	7	2	3	3
P103	52	2	2	4	5	4	3	2	2	4	3	2	1
P104	28	2	3	1	4	3	2	3	4	3	1	4	1

Database Construction and Entity Relationship Diagram

- We wanted a locally hosted database and our dataset is relatively small, so we chose to use a SQLite database.
- A simple way to create a SQLite database is to use the sqlite3 module in python.
- After creating a database, queries can be run with SQL syntax.
- A table was created for both the symptoms and the lifestyle choices.
- A combined table was created using a full outer join of the lifestyle and symptoms 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

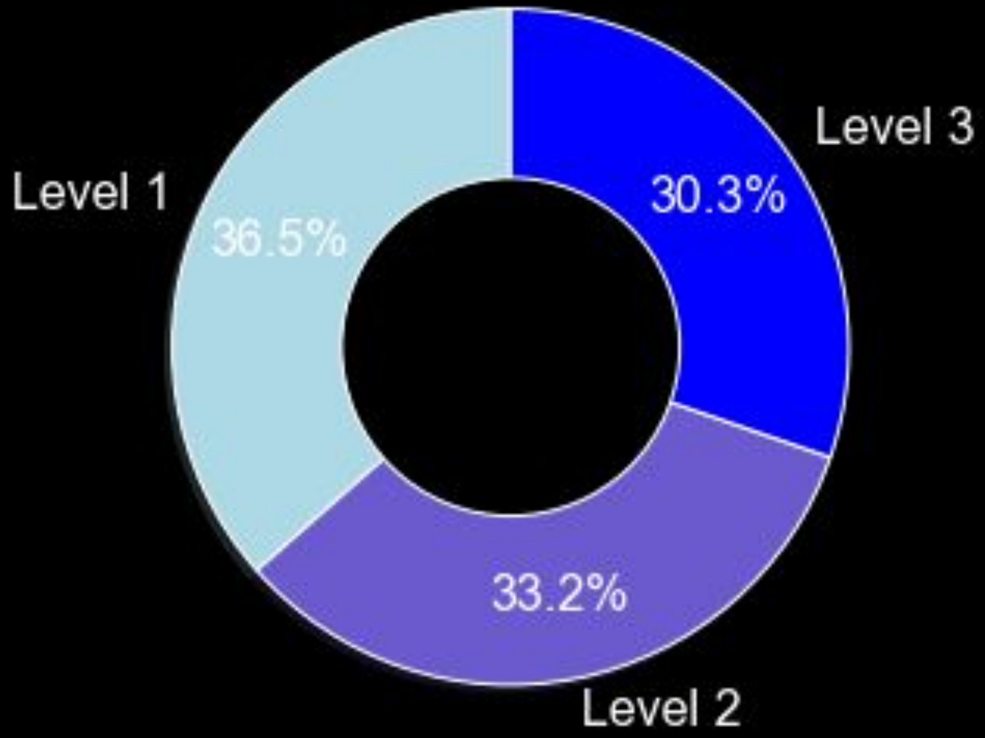
Combined	
Patient_ID	string
Age	int
Gender	int
Air_Pollution	int
Alcohol_Use	int
Dust_Allergy	int
Occupational_Hazards	int
Genetic_Risk	int
Chronic_Lung_Disease	int
Balanced_Diet	int
Obesity	int
Smoking	int
Passive_Smoking	int
Chest_Pain	int
Coughing_of_Blood	int
Fatigue	int
Weight_Loss	int
Shortness_of_Breath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

Entity relationship diagram for our database

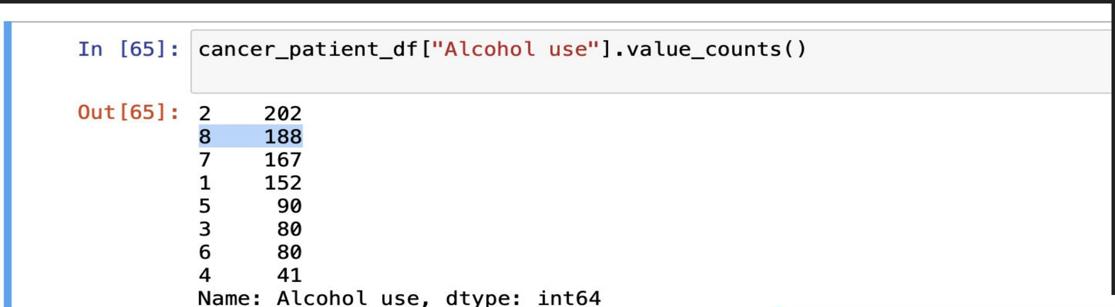
After looking over this data, we wanted to answer these three questions.

1. Which lifestyle choices are most associated with a higher severity of cancer?
2. How can we most accurately predict cancer severity using machine learning?
3. Which machine learning model predicts the severity of cancer most accurately?

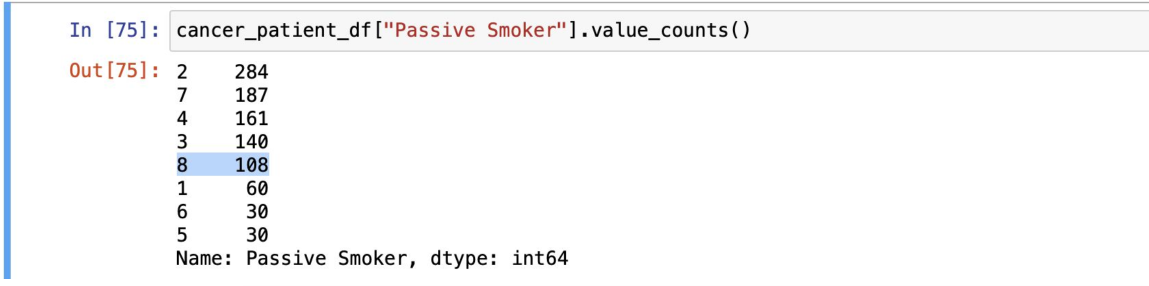
Cancer Severity by Level



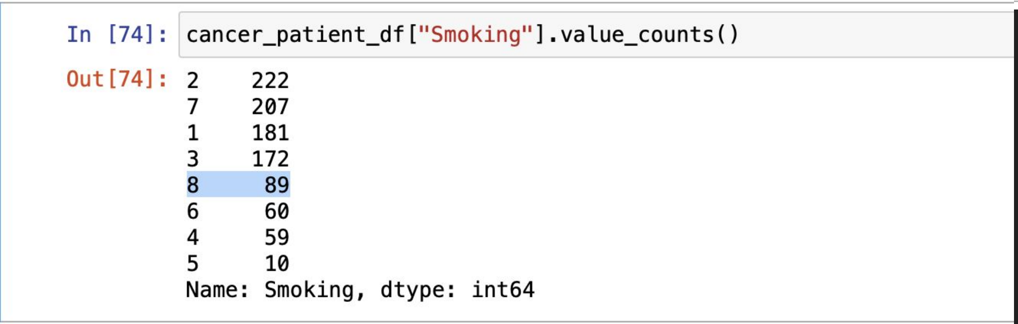
Top Three Columns with the Highest usage value (8)



```
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
```

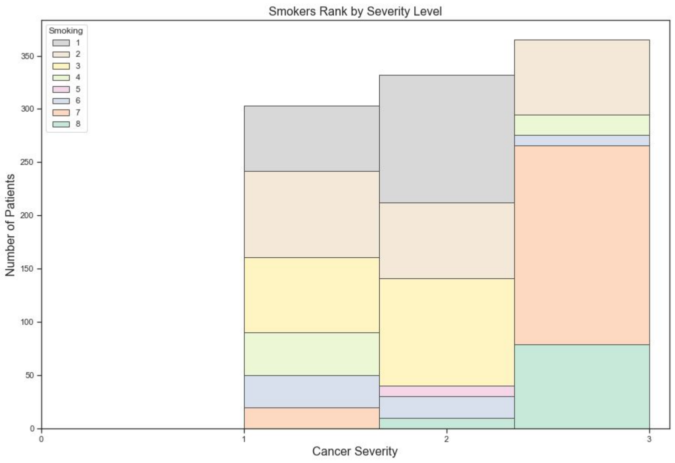


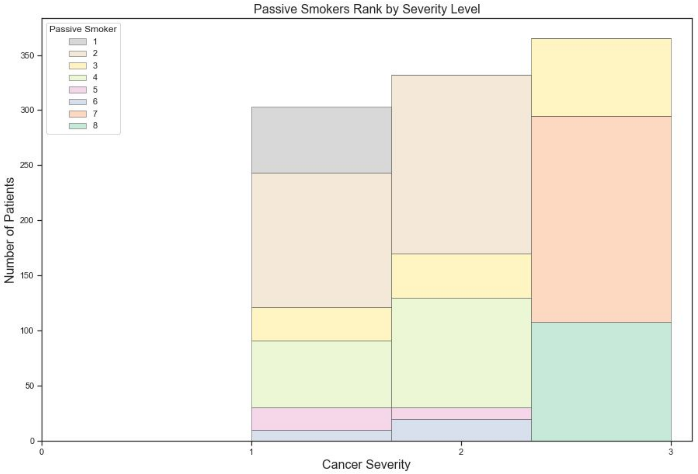
```
In [75]: cancer_patient_df["Passive Smoker"].value_counts()

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

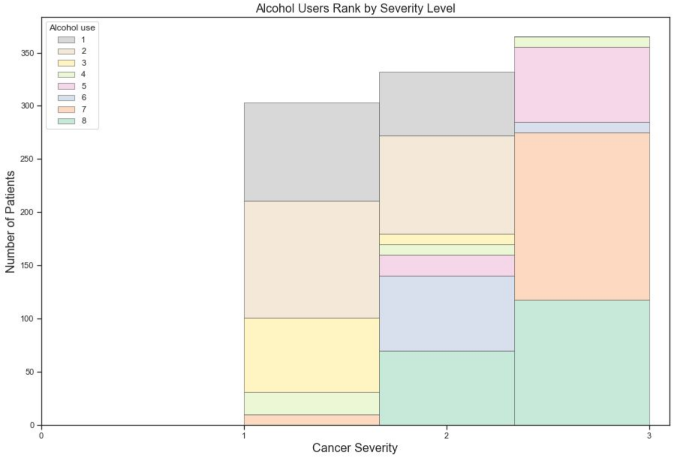


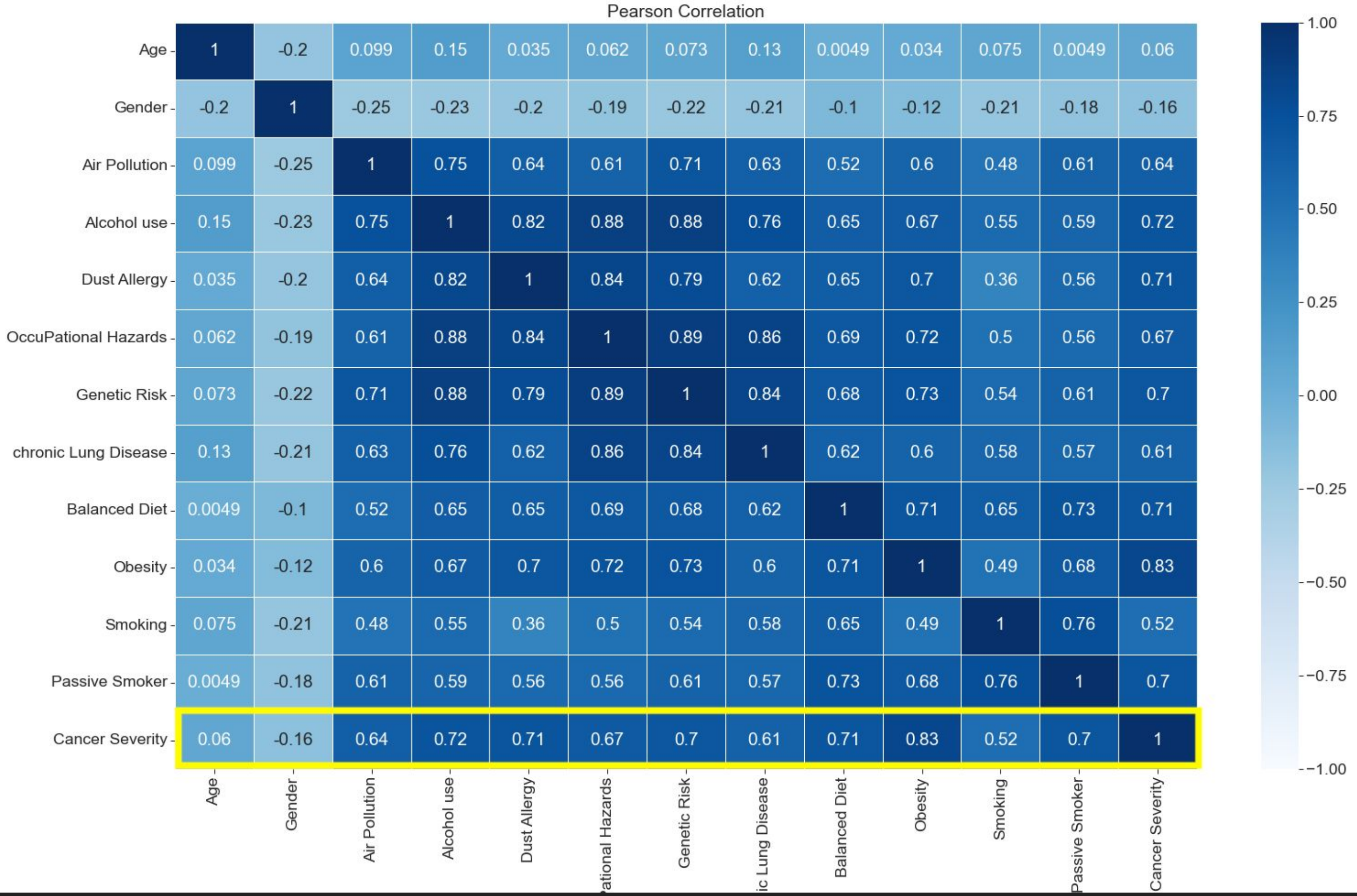
```
In [74]: cancer_patient_df["Smoking"].value_counts()

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

Cancer Severity Per Lifestyle Choice









	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	ational Hazards	Genetic Risk	ic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Cancer Severity
Age	1	-0.2	0.099	0.15	0.035	0.062	0.073	0.13	0.0049	0.034	0.075	0.0049	0.06
Gender	-0.2	1	-0.25	-0.23	-0.2	-0.19	-0.22	-0.21	-0.1	-0.12	-0.21	-0.18	-0.16
Air Pollution	0.099	-0.25	1	0.75	0.64	0.61	0.71	0.63	0.52	0.6	0.48	0.61	0.64
Alcohol use	0.15	-0.23	0.75	1	0.82	0.88	0.88	0.76	0.65	0.67	0.55	0.59	0.72
Dust Allergy	0.035	-0.2	0.64	0.82	1	0.84	0.79	0.62	0.65	0.7	0.36	0.56	0.71
OccuPational Hazards	0.062	-0.19	0.61	0.88	0.84	1	0.89	0.86	0.69	0.72	0.5	0.56	0.67
Genetic Risk	0.073	-0.22	0.71	0.88	0.79	0.89	1	0.84	0.68	0.73	0.54	0.61	0.7
chronic Lung Disease	0.13	-0.21	0.63	0.76	0.62	0.86	0.84	1	0.62	0.6	0.58	0.57	0.61
Balanced Diet	0.0049	-0.1	0.52	0.65	0.65	0.69	0.68	0.62	1	0.71	0.65	0.73	0.71
Obesity	0.034	-0.12	0.6	0.67	0.7	0.72	0.73	0.6	0.71	1	0.49	0.68	0.83
Smoking	0.075	-0.21	0.48	0.55	0.36	0.5	0.54	0.58	0.65	0.49	1	0.76	0.52
Passive Smoker	0.0049	-0.18	0.61	0.59	0.56	0.56	0.61	0.57	0.73	0.68	0.76	1	0.7
Cancer Severity	0.06	-0.16	0.64	0.72	0.71	0.67	0.7	0.61	0.71	0.83	0.52	0.7	1

Model Overview

- Model of Choice : Logistic Regression & SVM for the predicting the severity of cancer 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.
- 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 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 regression.

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

A powerful classification algorithm for predicting classification problems. 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.

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

Comparison Pros/Cons of using various models

Benefits of Logistic regression

1. Solving classification problem
2. Works with already identified independent variable

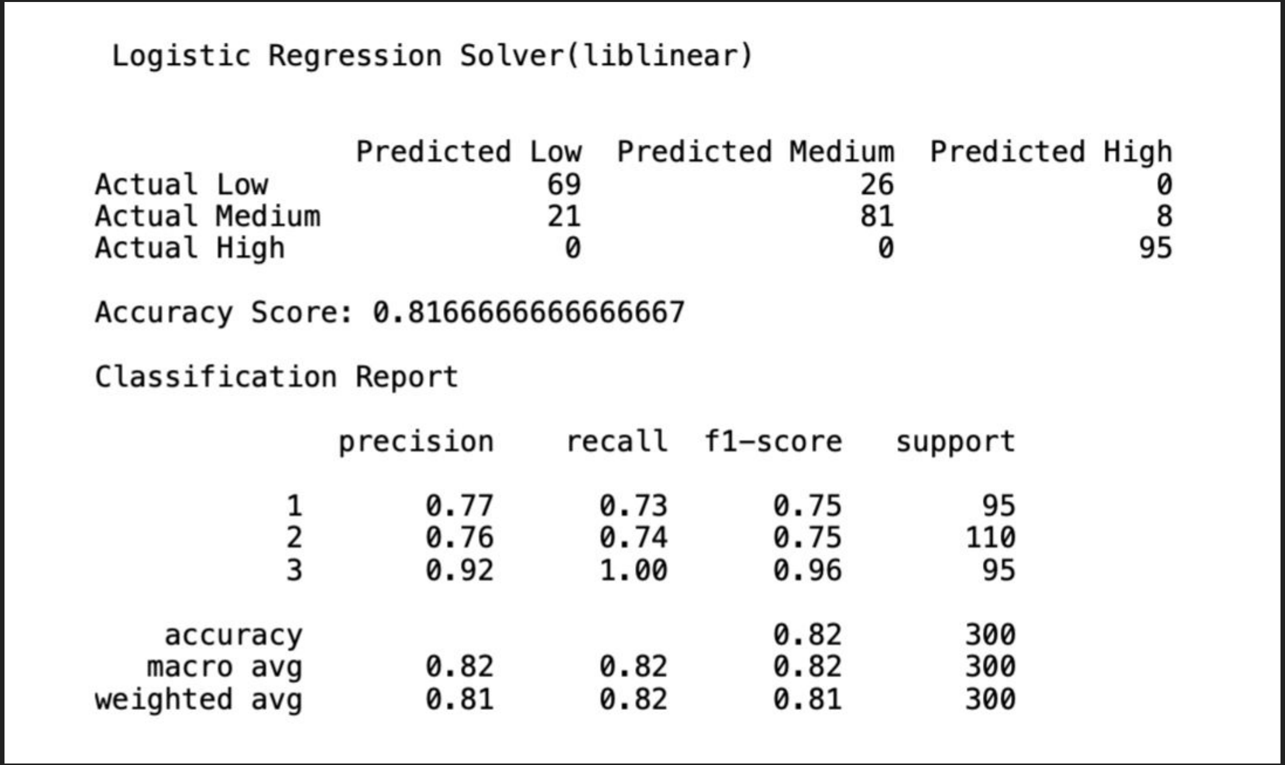
Cons

1. Vulnerable to overfitting
2. Can miss outliers / lower sensitivity for large unbalanced data sets

SVM

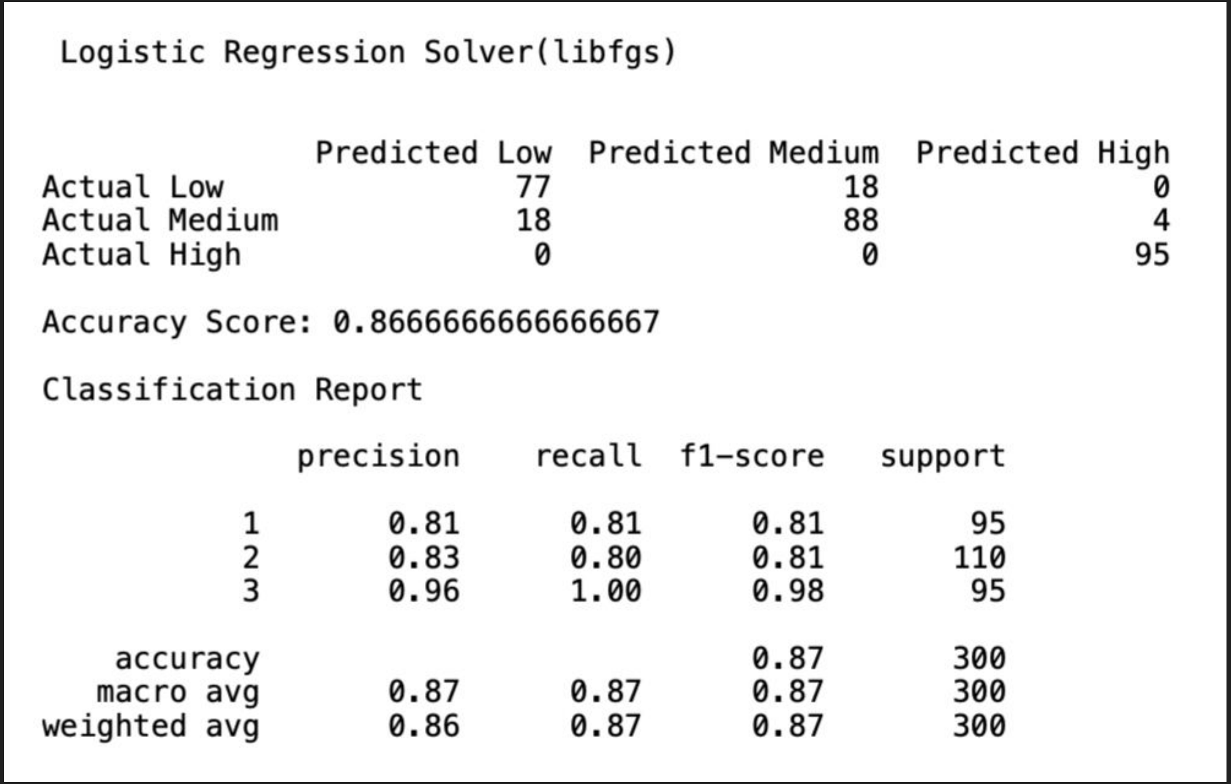
1. Tries to find the best margin that separates the classes that reduces the risk of error on the data
2. Risk of overfitting is less.
3. Gave us the best prediction for the model under study.

Logistic Regression (liblinear)



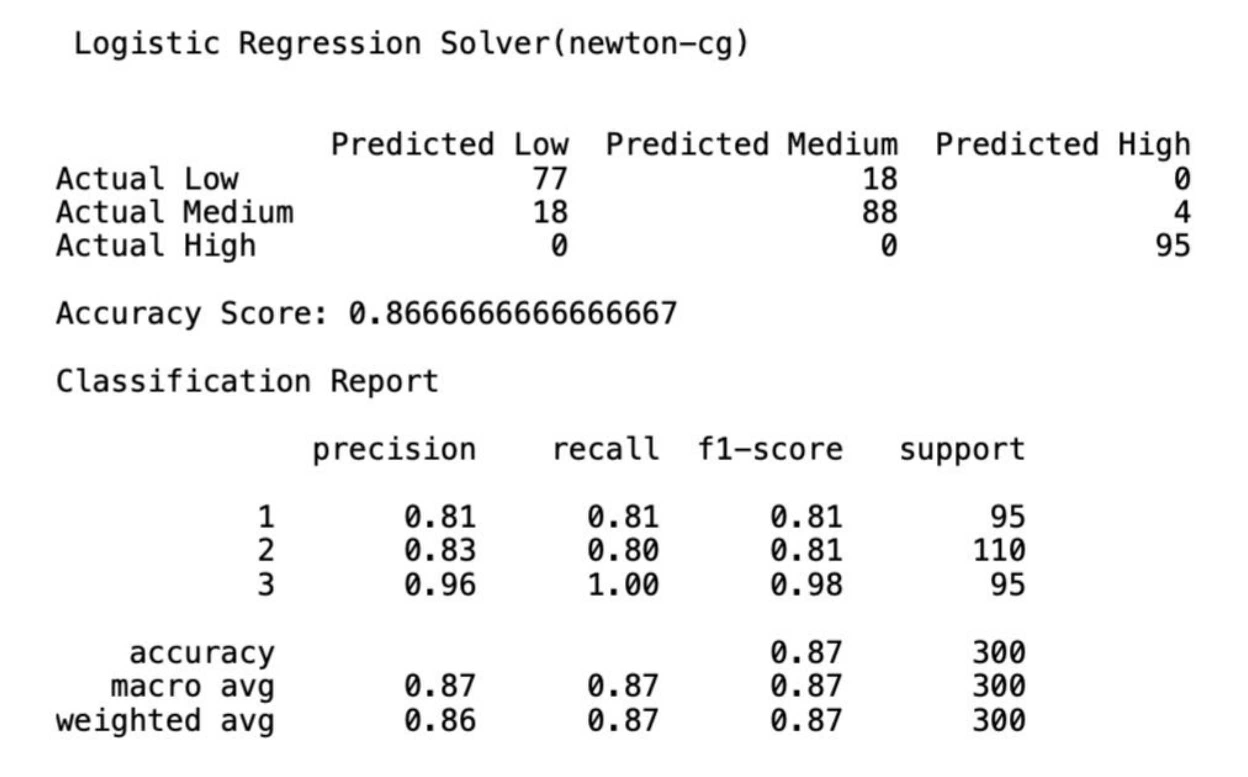
```
 Logistic Regression Solver(liblinear)


               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



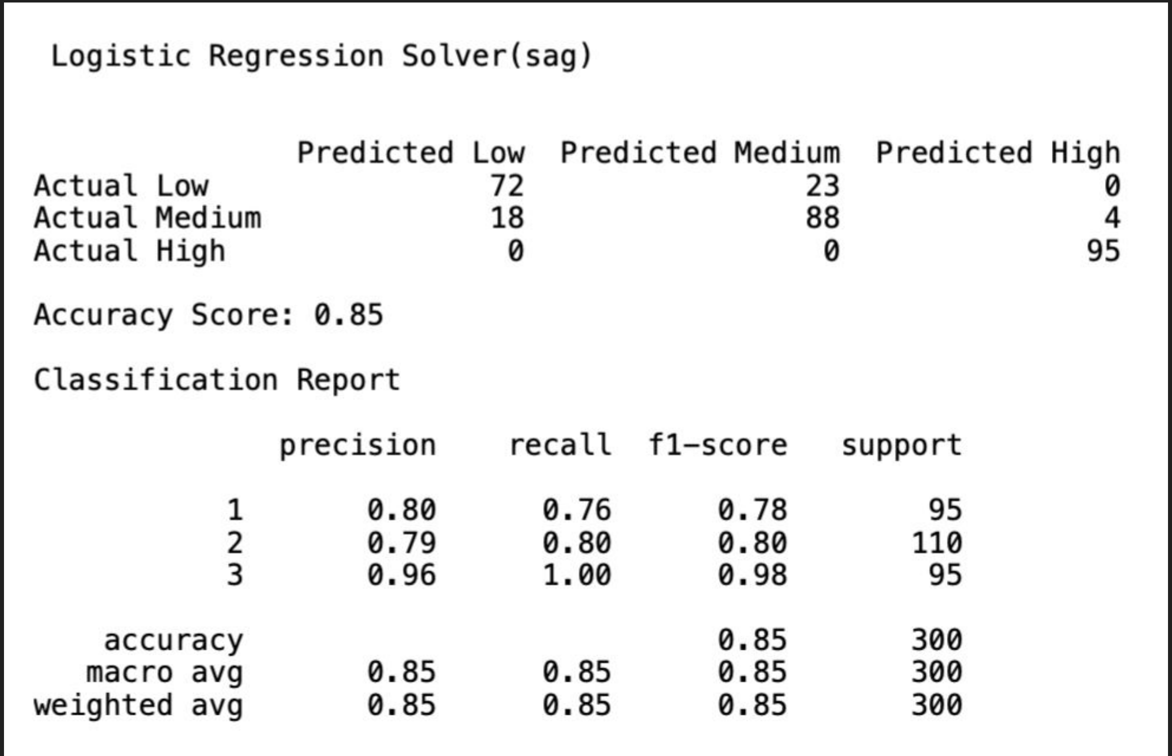
```
 Logistic Regression Solver(libfgs)


               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



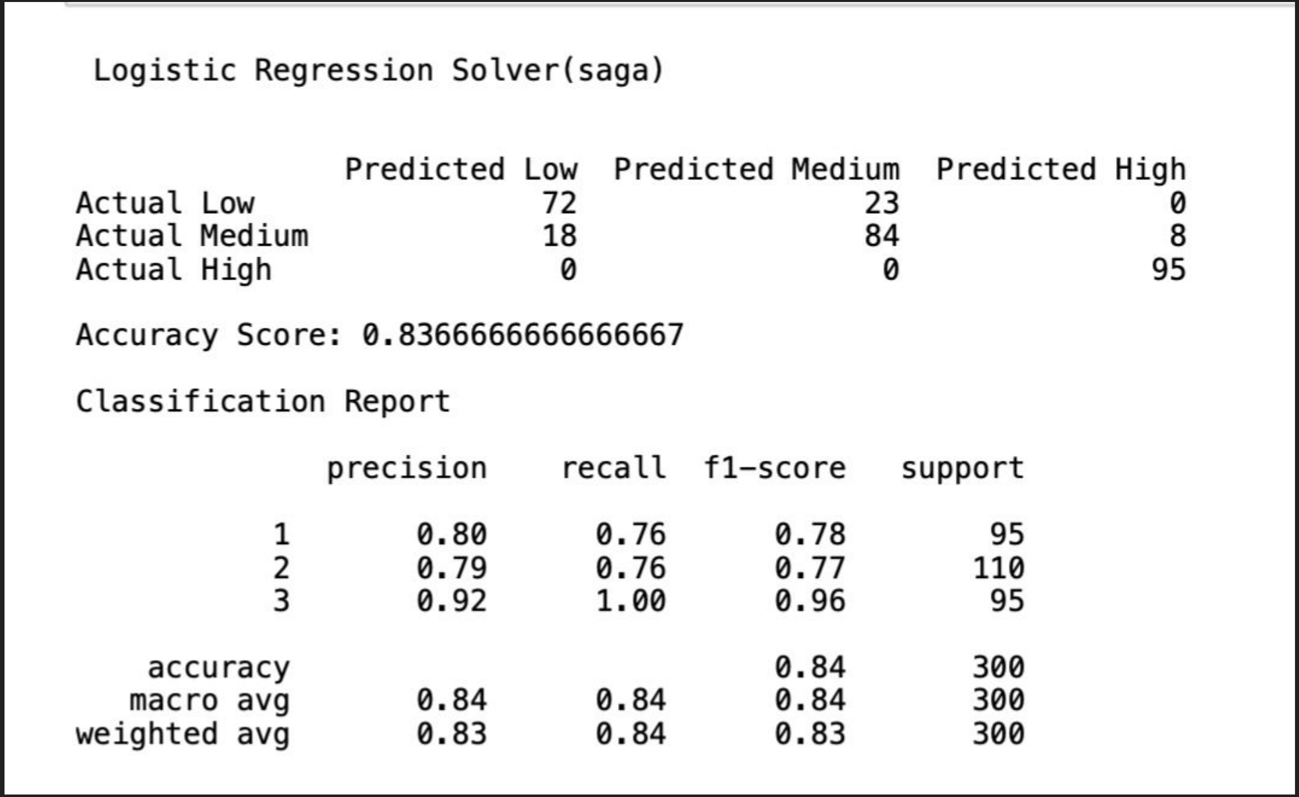
```
 Logistic Regression Solver(newton-cg)


              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



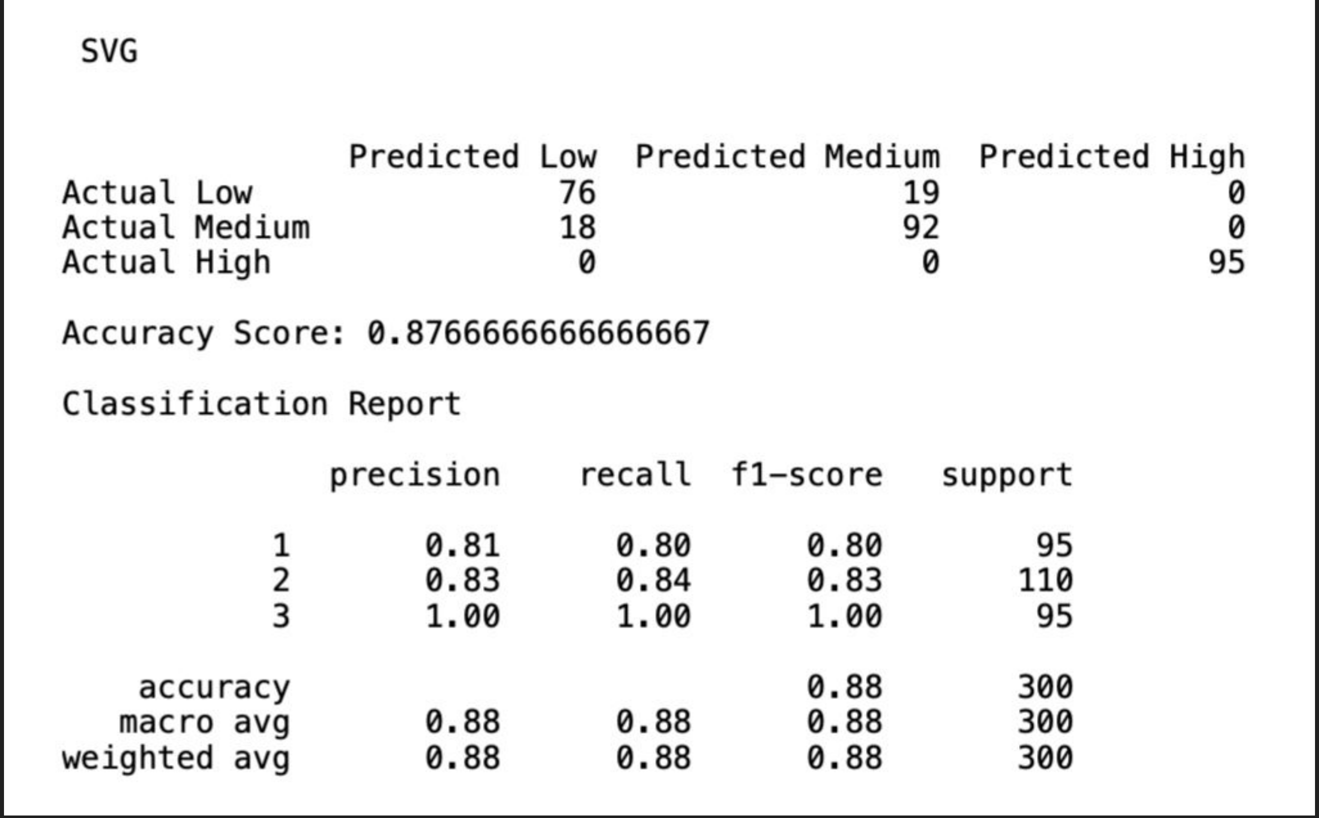
```
 Logistic Regression Solver(sag)


               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)

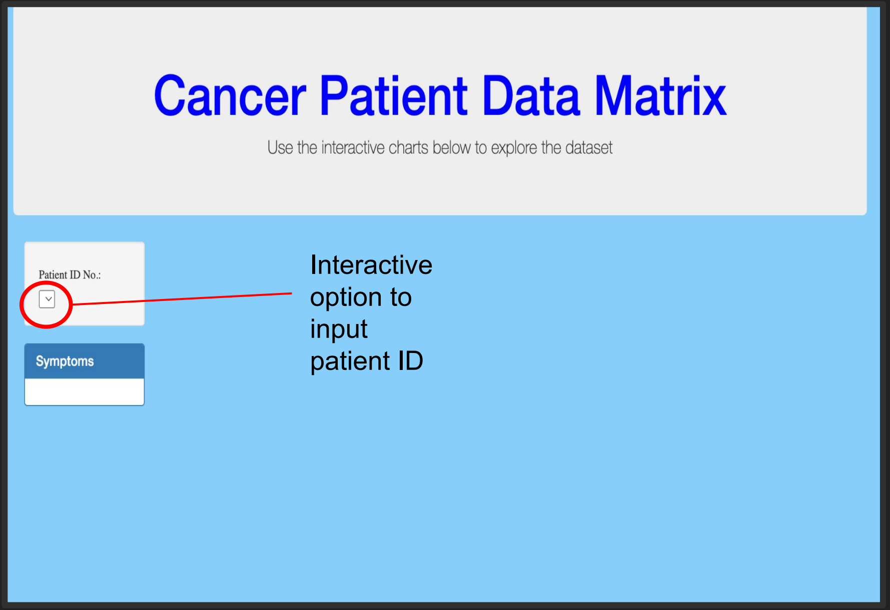


```
 SVG


               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

Early 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

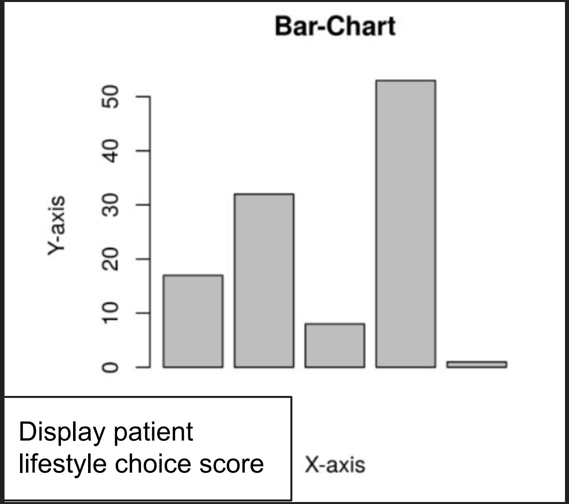
Use the interactive charts below to explore the dataset

Patient ID No.:

Symptoms

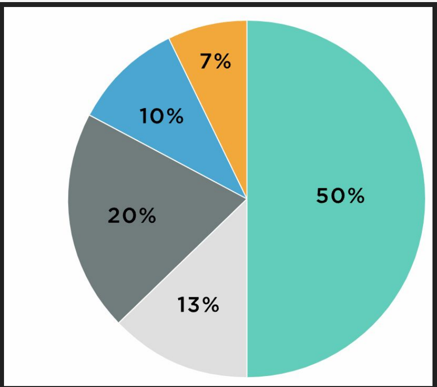
Interactive option to input patient ID

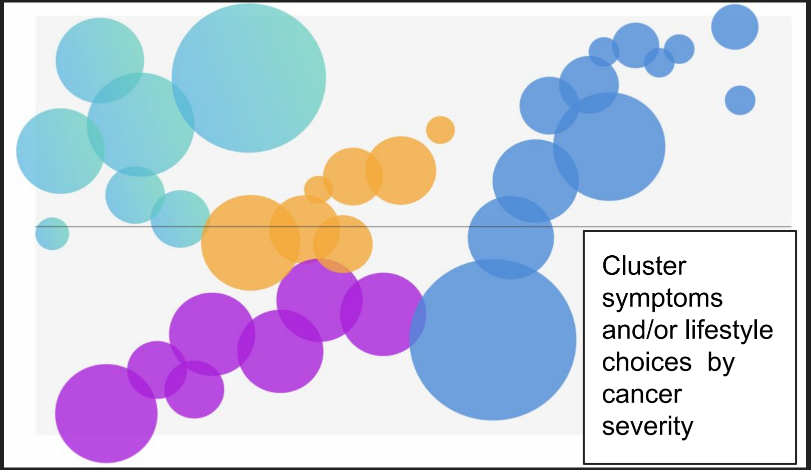
Blueprint



Display patient lifestyle choice score

Classify lifestyle scores at different cancer severity

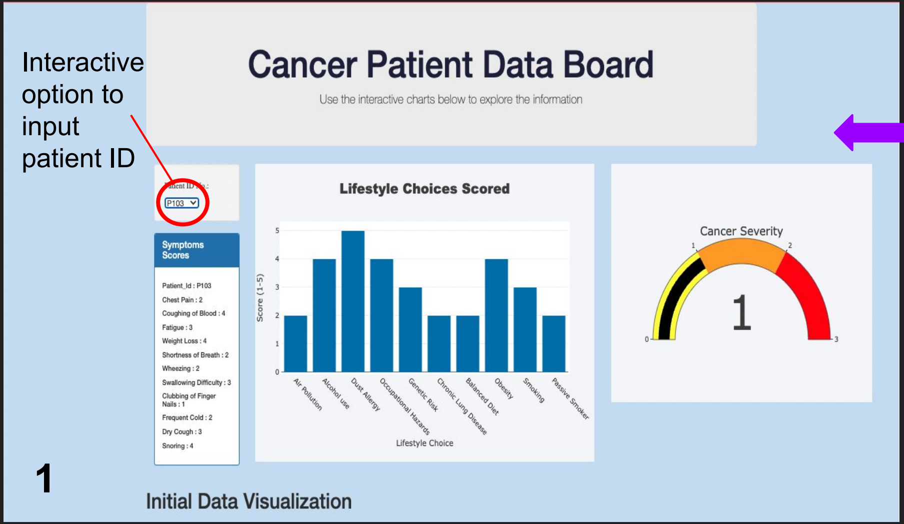




Cluster symptoms and/or lifestyle choices by cancer severity

Dashboard

Final View



Interactive option to input patient ID

Cancer Patient Data Board

Use the interactive charts below to explore the information

Patient ID No.:

P103

Symptoms Scores

Patient_Id : P103

Chest Pain : 2

Coughing of Blood : 4

Fatigue : 3

Weight Loss : 4

Shortness of Breath : 2

Wheezing : 2

Swallowing Difficulty : 3

Clubbing of Finger Nails : 1

Frequent Cold : 2

Dry Cough : 3

Snoring : 4

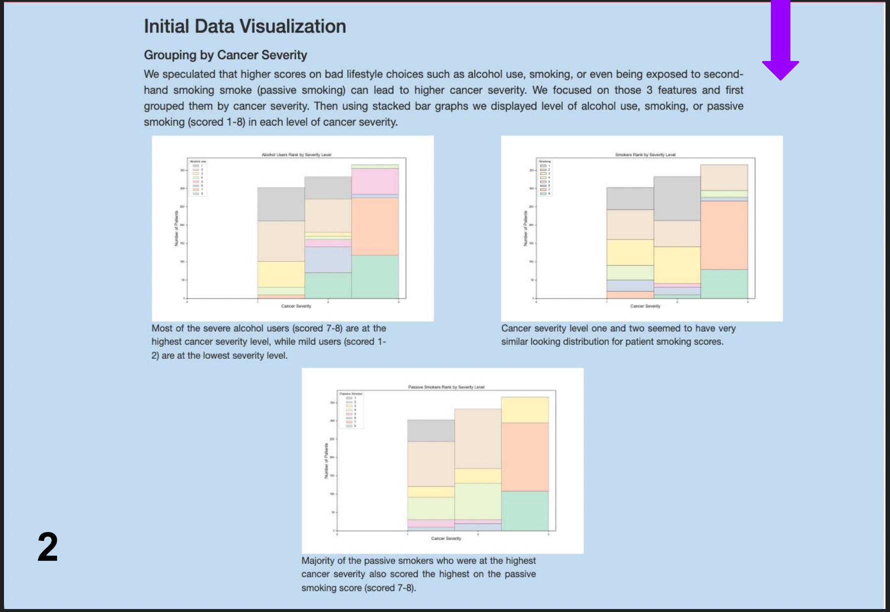
1

Initial Data Visualization

When selected, displays:

- Symptoms scores
- Lifestyle scores bar graph
- Cancer severity gauge

Classifies 3 hypothesized lifestyle habits by different cancer severity



Initial Data Visualization

Grouping by Cancer Severity

We speculated that higher scores on bad lifestyle choices such as alcohol use, smoking, or even being exposed to second-hand smoking smoke (passive smoking) can lead to higher cancer severity. We focused on those 3 features and first grouped them by cancer severity. Then using stacked bar graphs we displayed level of alcohol use, smoking, or passive smoking (scored 1-8) in each level of cancer severity.

Most of the severe alcohol users (scored 7-8) are at the highest cancer severity level, while mild users (scored 1-2) are at the lowest severity level.

Cancer severity level one and two seemed to have very similar looking distribution for patient smoking scores.

Majority of the passive smokers who were at the highest cancer severity also scored the highest on the passive smoking score (scored 7-8).

2

Current View of the Dashboard

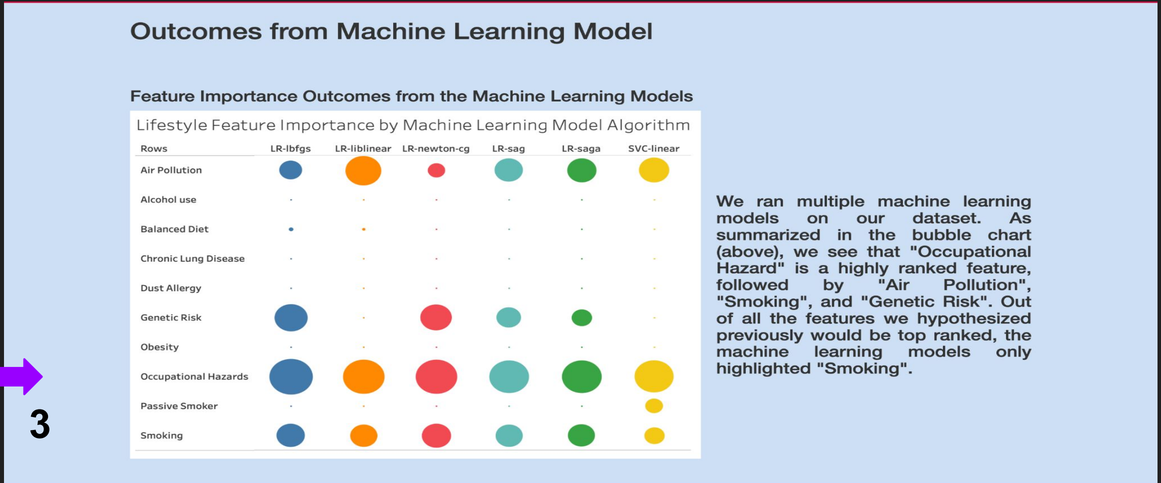
- Coded using **d3.json** with **CSS** styling and **Bootstrap** components
- Dashboard uses **HTML**

Dashboard

Final View

Bubble chart displays the **feature importance**

3



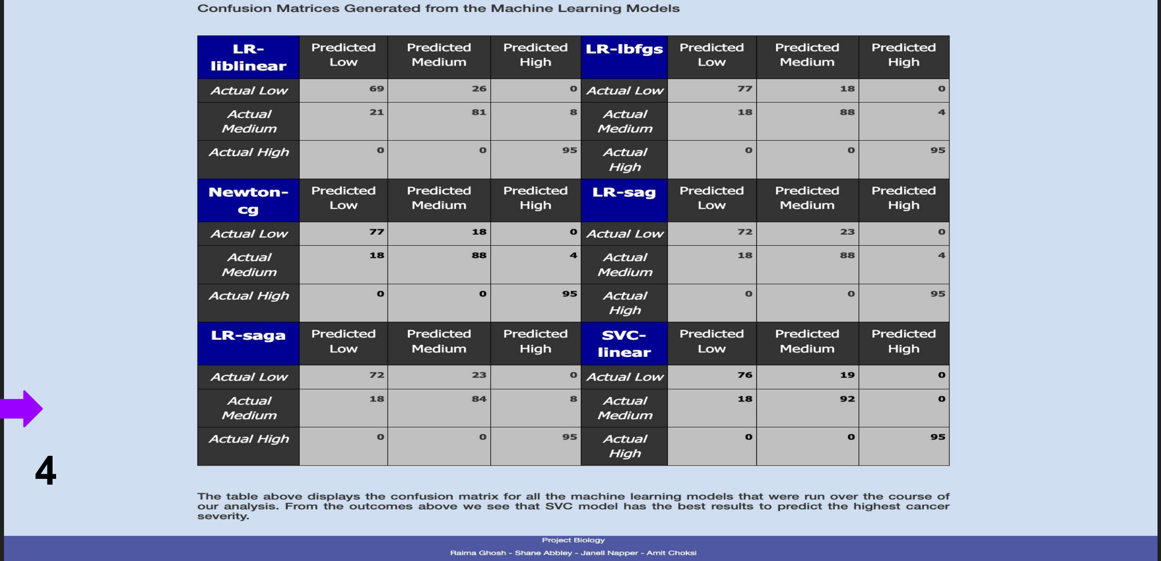
Outcomes from Machine Learning Model

Feature Importance Outcomes from the Machine Learning Models

We ran multiple machine learning models on our dataset. As summarized in the bubble chart (above), we see that "Occupational Hazard" is a highly ranked feature, followed by "Air Pollution", "Smoking", and "Genetic Risk". Out of all the features we hypothesized previously would be top ranked, the machine learning models only highlighted "Smoking".

Confusion matrices from the 6 ML models displayed as table

4



Confusion Matrices Generated from the Machine Learning Models

LR-liblinear	Predicted Low	Predicted Medium	Predicted High	LR-lbfgs	Predicted Low	Predicted Medium	Predicted High
Actual Low	69	26	0	*Actual Low*	77	18	0
Actual Medium	21	81	8	*Actual Medium*	18	88	4
Actual High	0	0	95	*Actual High*	0	0	95
Newton-cg	Predicted Low	Predicted Medium	Predicted High	**LR-sag**	Predicted Low	Predicted Medium	Predicted High
Actual Low	77	18	0	*Actual Low*	72	23	0
Actual Medium	18	88	4	*Actual Medium*	18	88	4
Actual High	0	0	95	*Actual High*	0	0	95
LR-saga	Predicted Low	Predicted Medium	Predicted High	**SVC-linear**	Predicted Low	Predicted Medium	Predicted High
Actual Low	72	23	0	*Actual Low*	76	19	0
Actual Medium	18	84	8	*Actual Medium*	18	92	0
Actual High	0	0	95	*Actual High*	0	0	95

The table above displays the confusion matrix for all the machine learning models that were run over the course of our analysis. From the outcomes above we see that SVC model has the best results to predict the highest cancer severity.

Project Biology

Raima Ghosh - Shane Abbley - Janell Napper - Amit Choksi

Dashboard

Final View- Demo

After our experiment, we answered all three of these questions.

1. Which lifestyle choices are most associated with a higher severity of cancer?
   Our model predicted occupational hazards, air pollution, and smoking as the three most relevant lifestyle factors.

2. How can we most accurately predict cancer severity using machine learning?
   We can try to run the algorithm again without some of the features that had a minimal impact on this models predictions.

3. Which machine learning model predicts the severity of cancer most accurately?
   We found that support vector machine (SVM) gave us the most accurate results.

Questions?