Predicting Lung Cancer Risk Levels at Home

By Julian Cheng

Table of Contents

- Methodology

- Processing

- Results

- Conclusion

- Next Steps

- Appendix

The Problem

Is it possible to predict lung cancer risk levels in human patients using self-reported statistics?

Answer

I developed a classification model that can accurately predict a patient's risk of lung cancer within three broad bins using self-reported statistics.
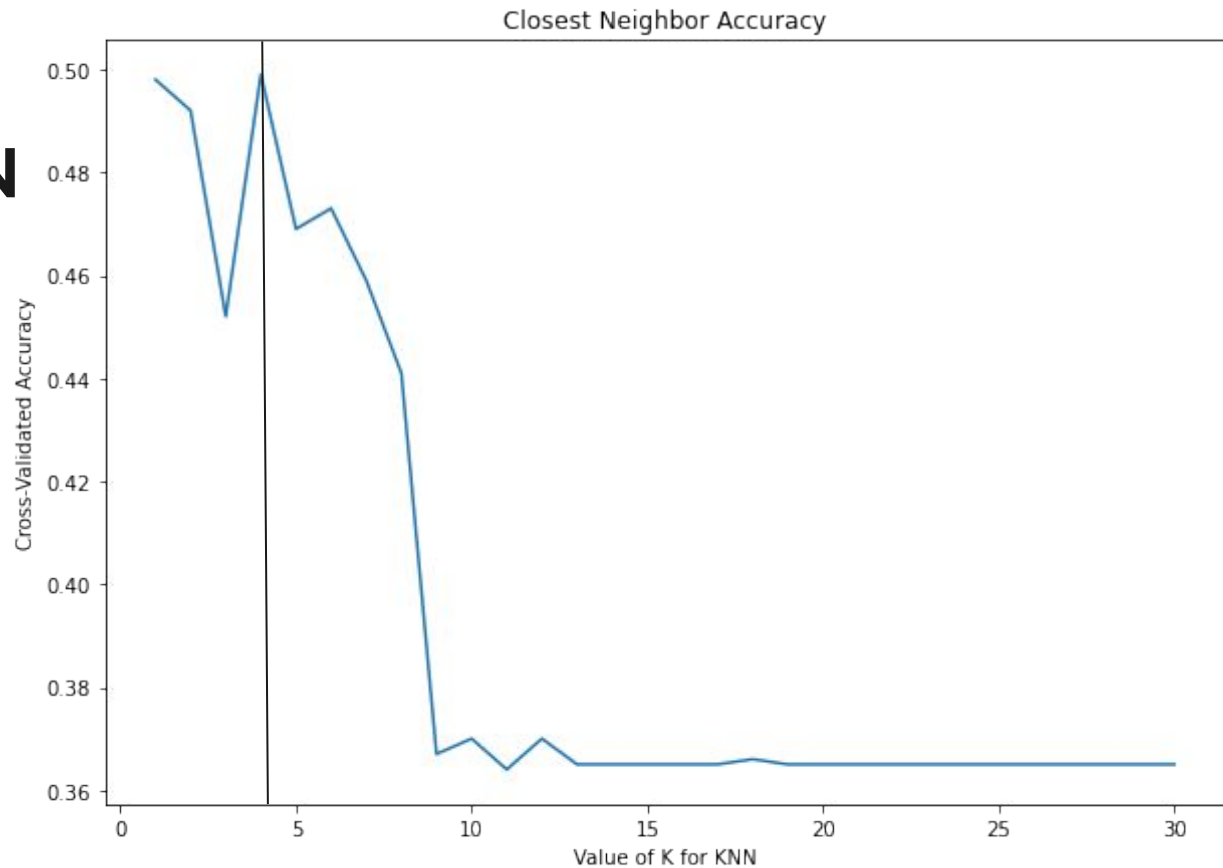
Methodology - Data Balancing

Risk Level:	Count:
Low	303
Medium	332
High	365

Methodology - KNN

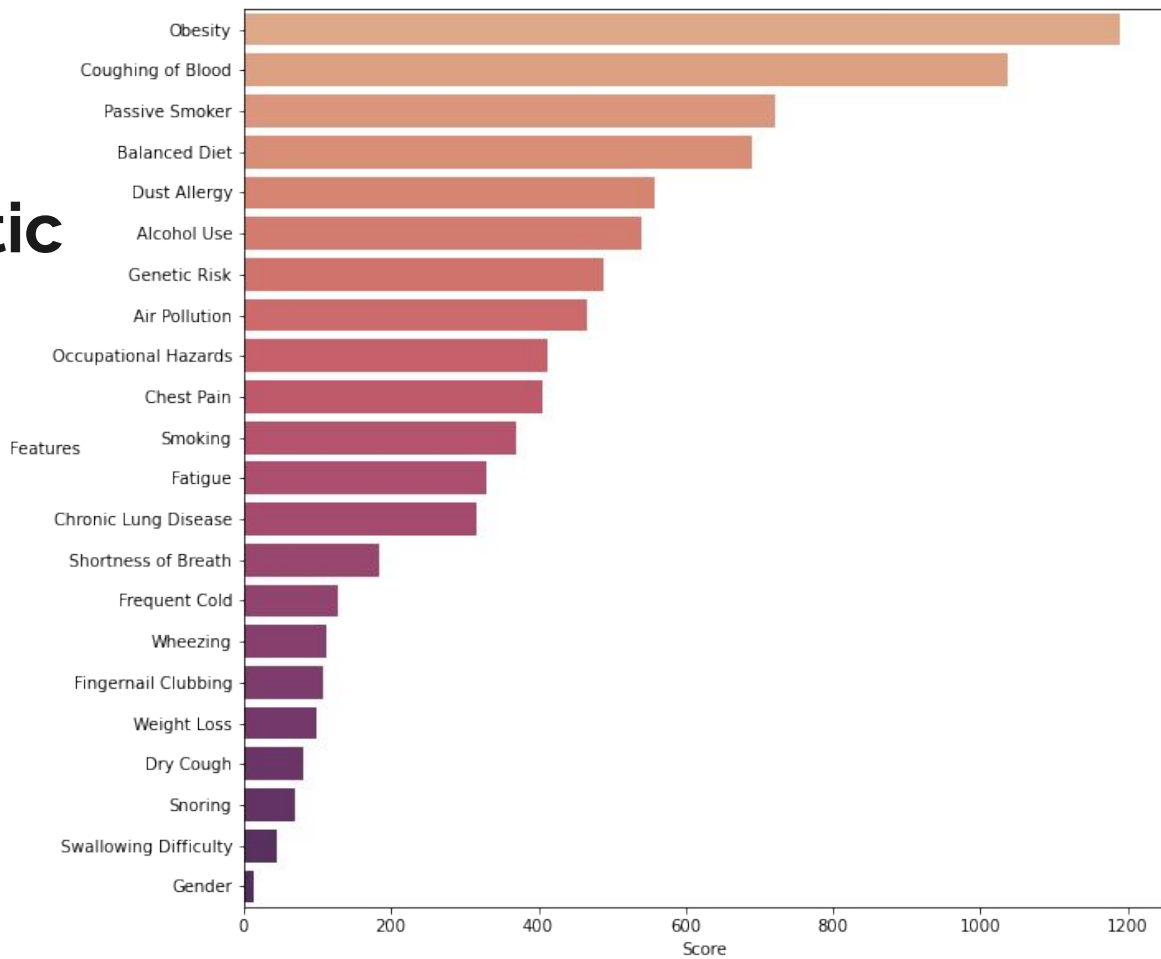GridSearchCV:

Optimal K value = 4
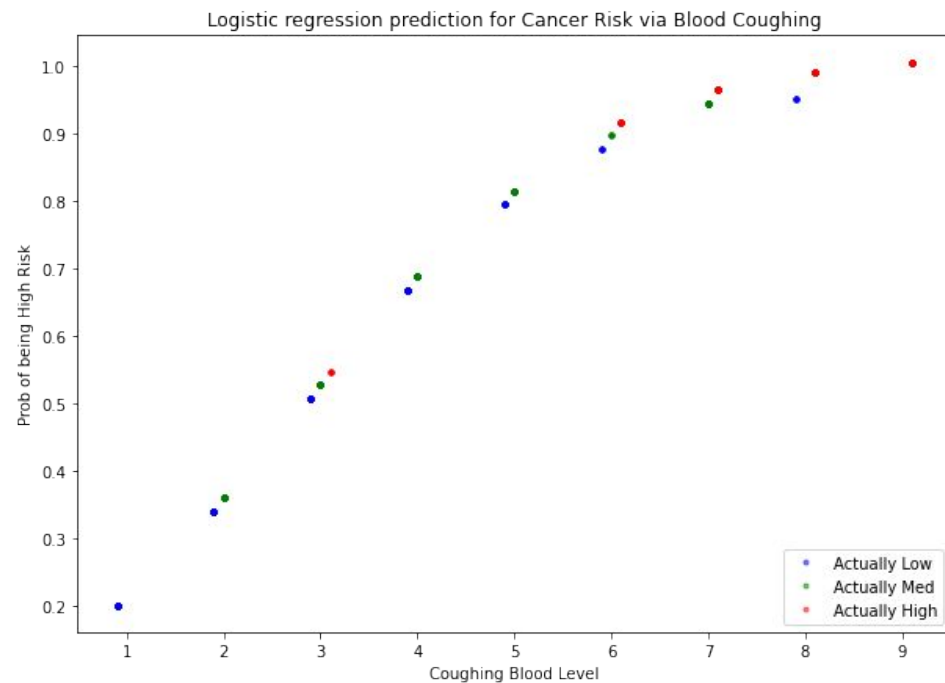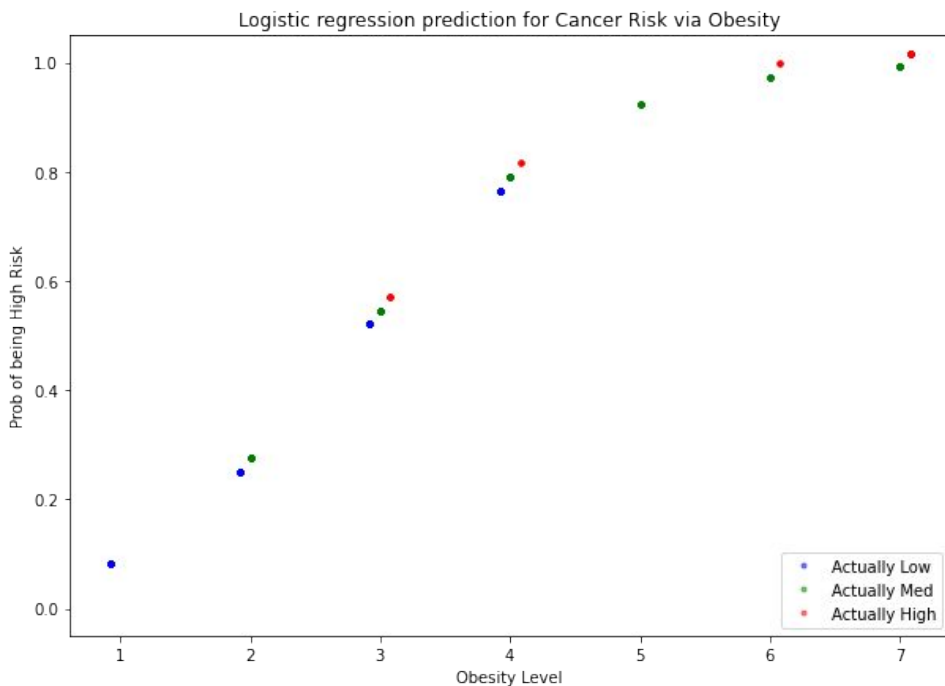

Closest Neighbor Accuracy

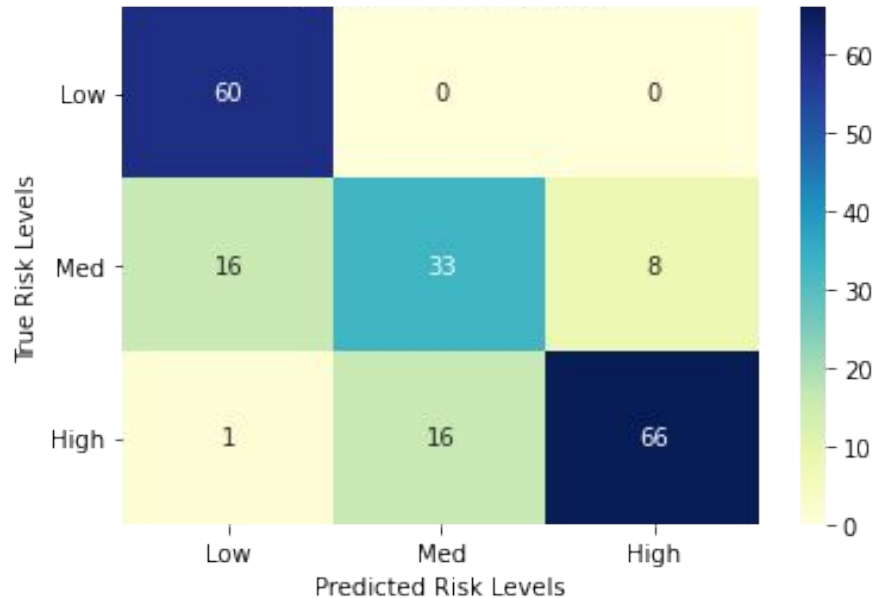# Methodology - Logistic

Logistic Models:

- Obesity
- Coughing of Blood

Methodology - Logistic Graphs

Logistic regression prediction for Cancer Risk via Obesity
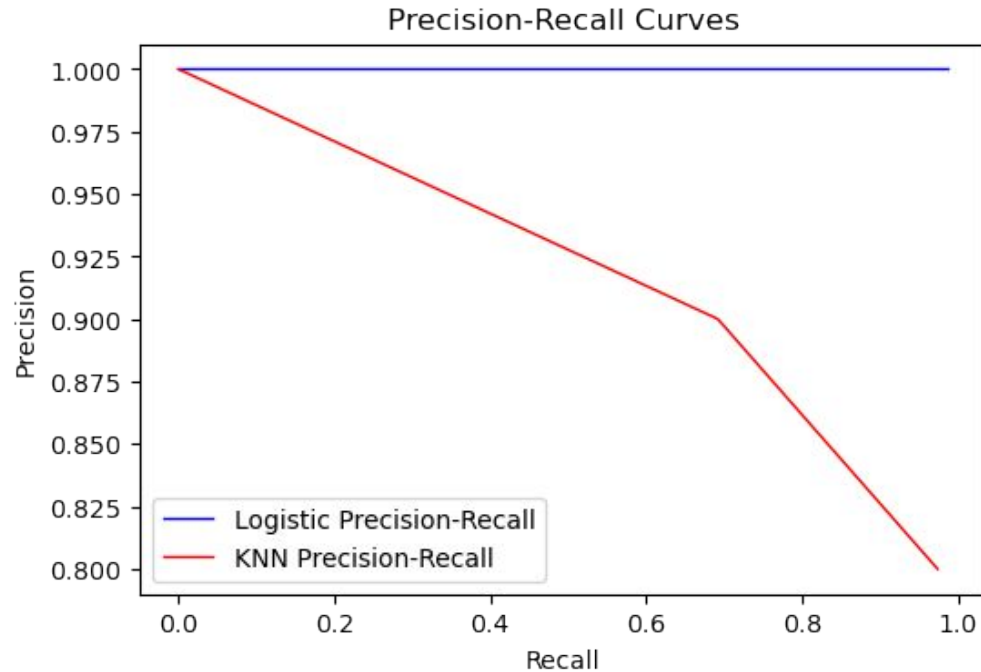
Logistic regression prediction for Cancer Risk via Blood Coughing

Methodology - Confusion Matrices
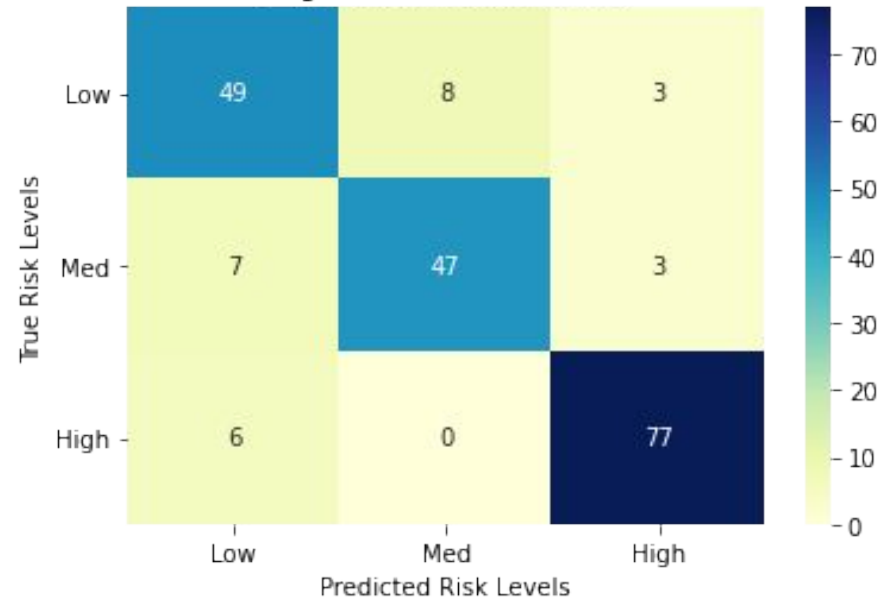
Results - Precision Recall

Methodology - Model Scores

X	Accuracy		Precision (per class)			Recall (per class)			F1 Score (per class)		
Model	Train	Test	Low	Med	High	Low	Med	High	Low	Med	High
KNN	**0.93**	0.80	0.78	0.67	0.89	**1.00**	0.58	0.80	**0.88**	0.62	0.84
Log	0.90	**0.87**	**0.79**	**0.85**	**0.93**	0.82	**0.82**	**0.93**	0.80	**0.84**	**0.93**

Results - Binarized ROC

ROC curve for KNN classification Cancer Risk

ROC curve for Logistic classification Cancer Risk

True positive rate

False positive rate

Low ROC Score = 0.98
Med ROC Score = 0.83
High ROC Score = 0.94

Low ROC Score = 1.00
Med ROC Score = 0.97
High ROC Score = 1.00
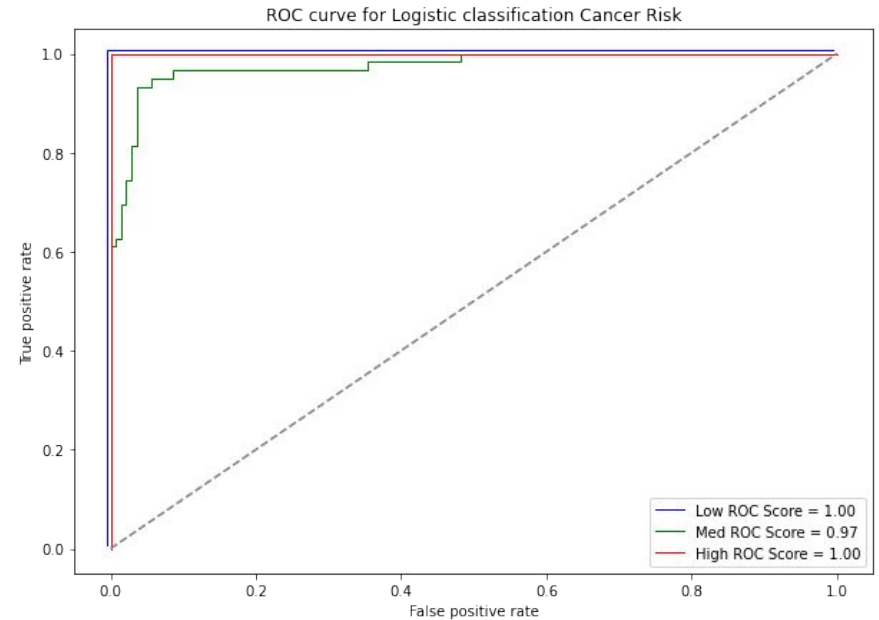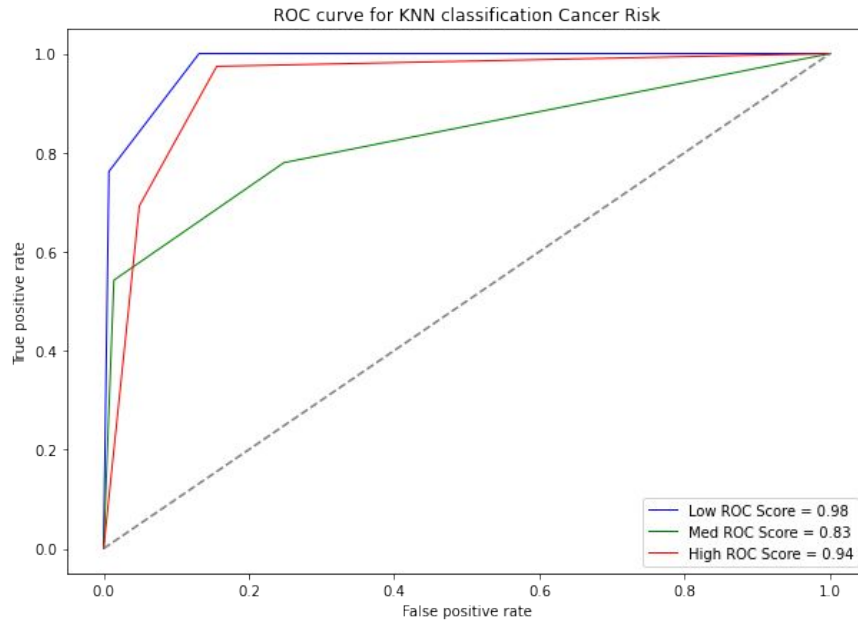
Conclusion - Precision vs. Recall

Precision - Calculating false negatives

Recall - Calculating false positives

A higher recall will catch more positive cases

A decreasing threshold will increase the recall at the cost of precision

A lower precision will result in more negatives being tested
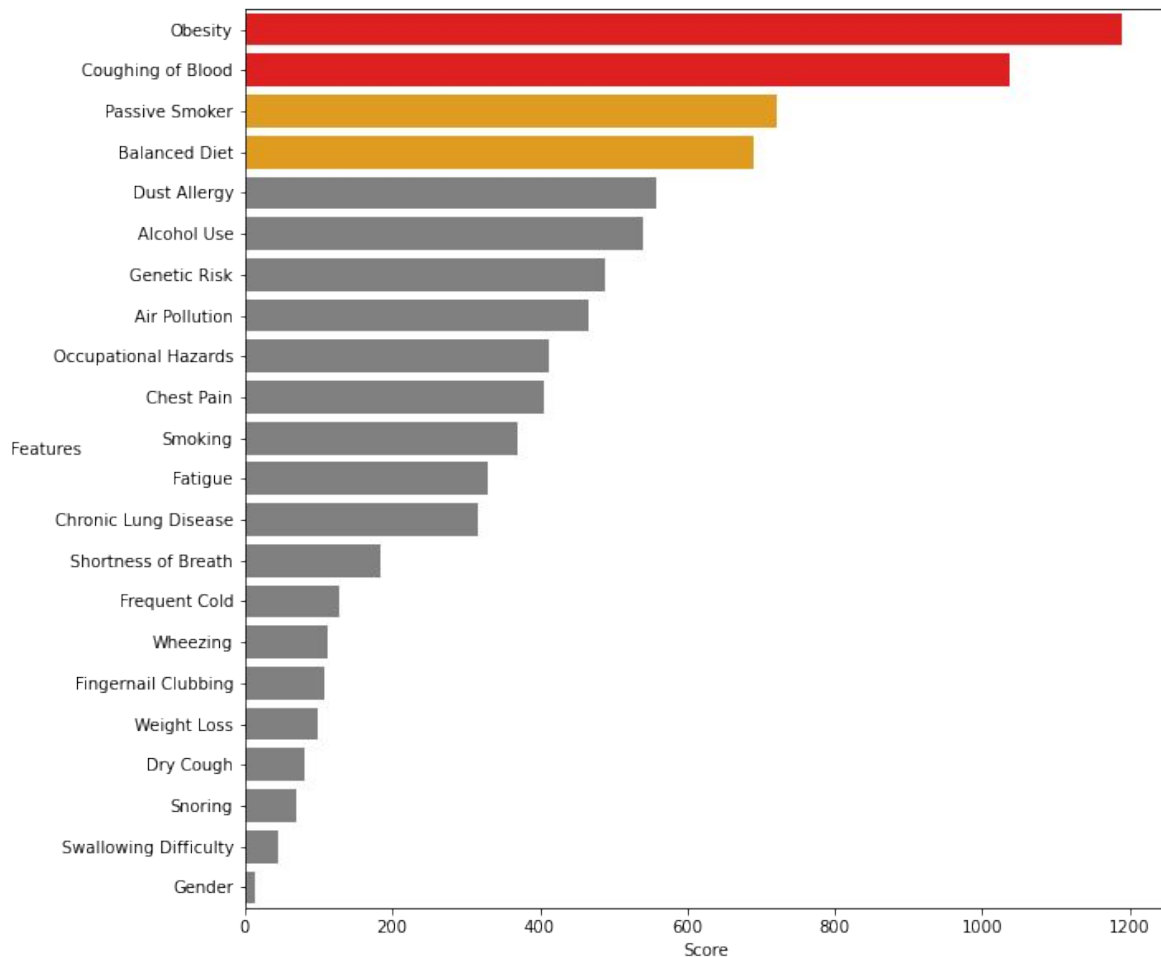
Conclusion - Flask

Balanced Diet (1-7)

Obesity Level (1-7)

Blood Coughing (1-9)

Exposure to Secondhand Smoke (1-8)

Submit

prediction:

Next Steps

- Kaggle - cleaned, balanced, suspiciously accurate
- More descriptive features and classes
- More visuals!

Appendix

https://www.kaggle.com/christopherwsmith/how-to-predict-lung-cancer-levels-100-accuracy