Yusuf Karadogan

Final Report:
Lung Cancer Prediction

Problem Statement

Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale.

Lung cancer is the leading cause of cancer deaths worldwide.

People who smoke have the greatest risk of lung cancer, though lung cancer can also occur in people who have never smoked. The risk of lung cancer increases with the length of time and number of cigarettes you've smoked. If you quit smoking, even after smoking for many years, you can significantly reduce your chances of developing lung cancer.

Lung cancer typically doesn't cause signs and symptoms in its earliest stages. Signs and symptoms of lung cancer typically occur when the disease is advanced. [Source of information: MayoClinic]

Signs and symptoms of lung cancer may include:

- A new cough that doesn't go away
- Coughing up blood, even a small amount
- Shortness of breath
- Chest pain
- Hoarseness
- Losing weight without trying
- Bone pain
- Headache

Along with these signs and symptoms that are provided through MayoClinic our dataset also provides 16 more additional variables that might provide insight to predicting Lung Cancer diagnosis.

As the cancer progresses and continues to spread, just like other forms of cancers, it becomes harder to treat and the survival rate decreases significantly. By analyzing patients' lifestyle and symptoms, I'll try to predict the risk level of being diagnosed with lung cancer. The success of a patient being diagnosed as early as possible, would add years if not decades to a patient's life, with the current advancement of immuno and chemo therapies.

After analyzing and optimizing, I was able to compare over a dozen different machine learning models. When compared side by side the majority of the models were able to achieve an 100% accuracy, precision,and AUC score. This process can be repeated for many other startups and with more variables can be improved.

Data Wrangling

The initial data set contained 1000 rows and 24 columns, while it was not a huge set, it was still enough to work with and build upon. I started analysing the dataset as a whole, from there focused on finding if there were any missing values for each column. Luckily none of the columns had any missing values. Overall this was a well balanced dataset. The Level column, which was our target for prediction, was distributed evenly with the data set divided as below;

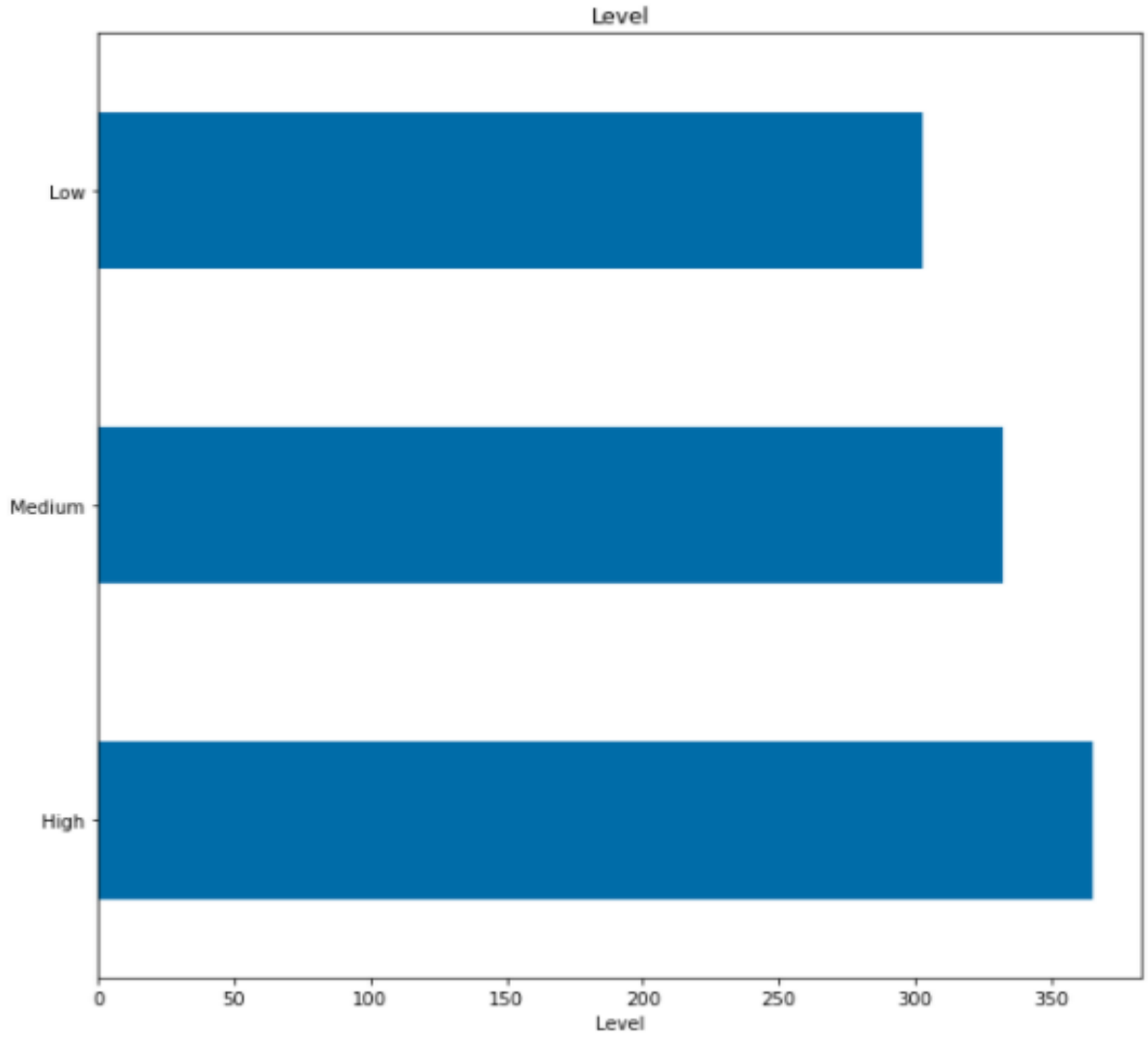
Level Distribution

65/1000 = 36.5% high level

332/1000 33.2% medium level

303/1000 = 30.3% low level

Next I looked to see if there were any duplicate rows, fortunately there were only were not any duplicated patients. The final shape of my dataset at this stage remained with 1000 rows and 24 columns.



Exploratory Data Analysis

Since there were not any missing values or duplicates, I did not have to to fill or drop rows for the exploratory data analysis. One thing that I did do to prep my data for future steps was to change the values of the Level column. I was able to replace "Low" with "1", "Medium" with "2", "High" with "3".

Next I wanted to visualize the different risk levels with its correlation to gender. Fig. 1 gender count and Fig. 2 displays the different genders with levels.

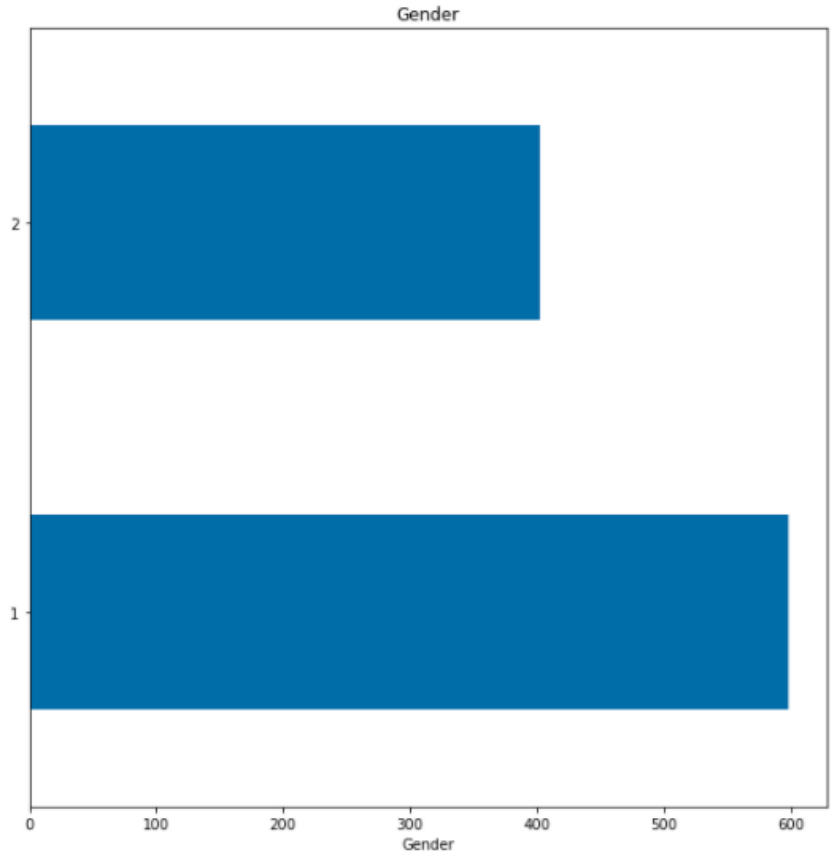


Figure 1: Gender Count

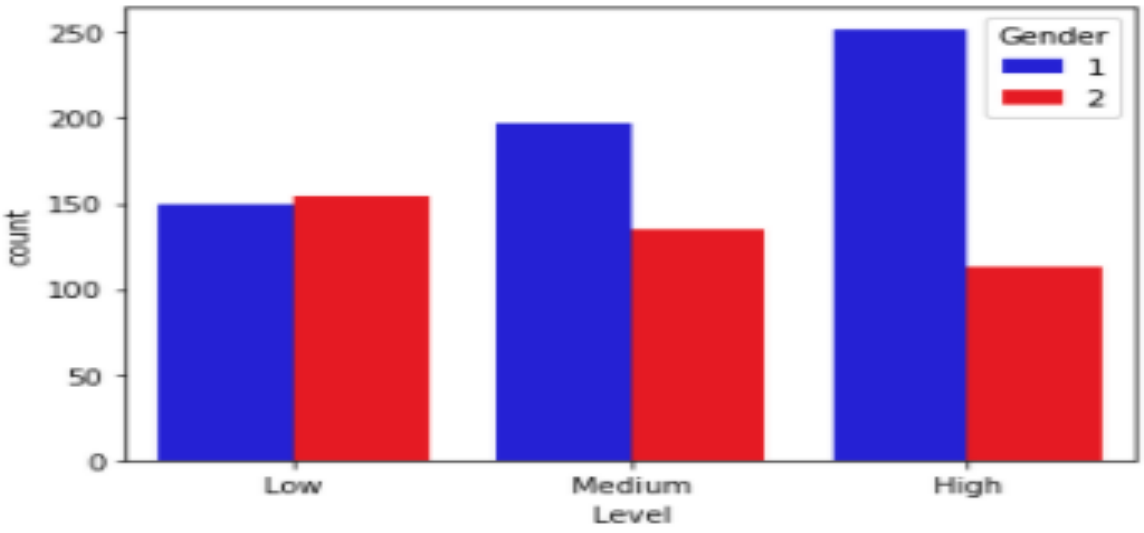


Figure 2: Bar Plot of top states in funding

I looked at the correlations between every column, which resulted in some strong correlations between the genetic risk and other columns These correlations between the time value columns ranged between 0.83 and 0.89, which showed significance. After some optimization and tuning, the Fig. 3 below shows some of these correlations.

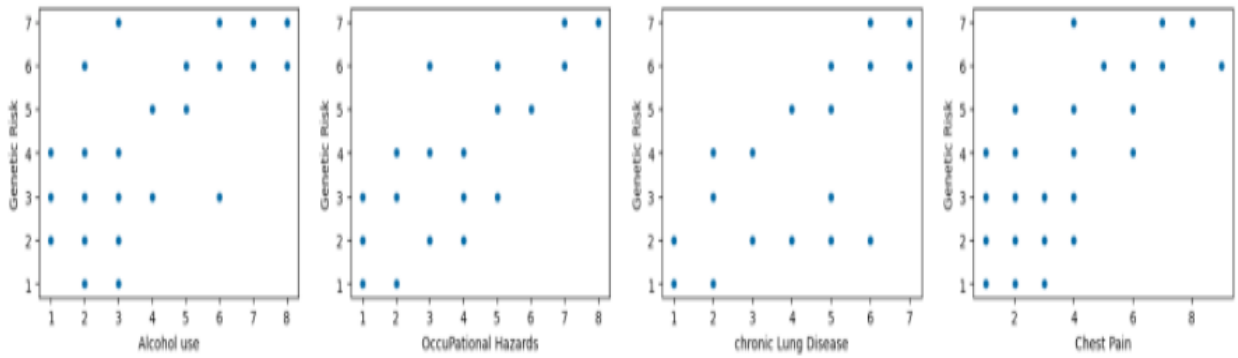
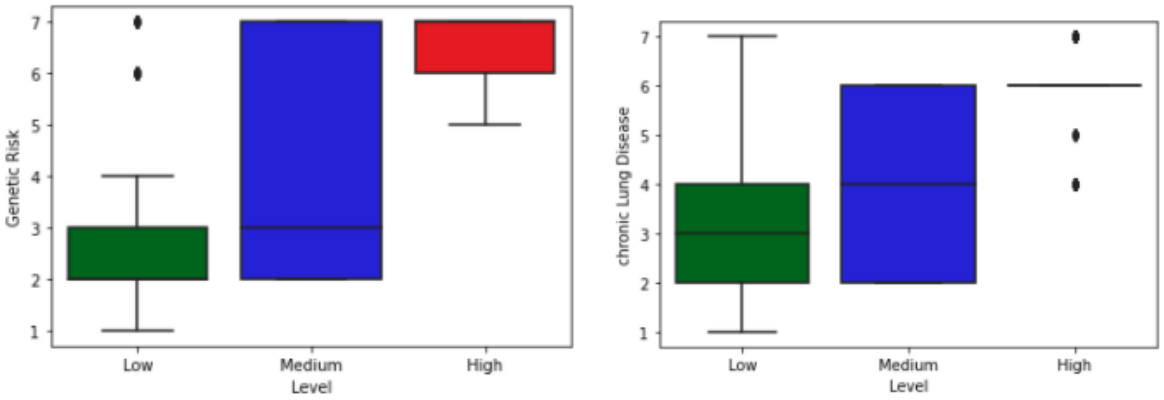


Figure 1: Scatter Plot of the genetis correlations

Next I looked at some of the outliers. Fig. 4 shows there weren't any outliers since they are within the limits of the assessment.



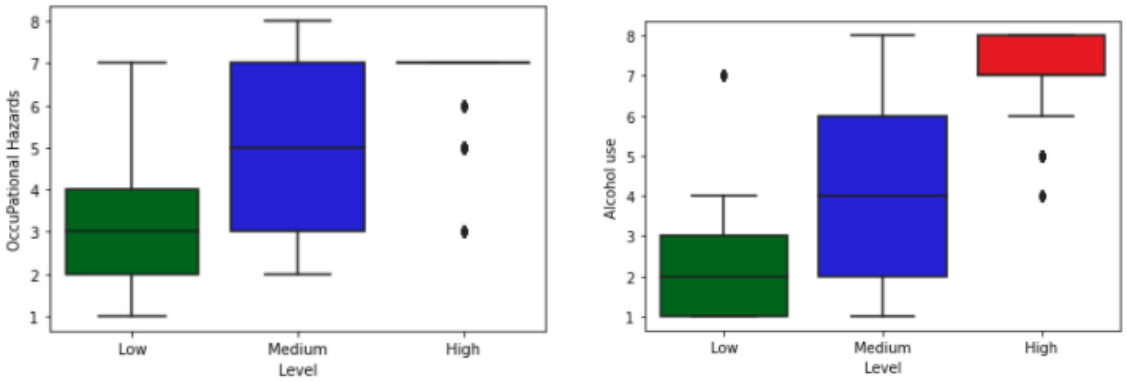


Figure 4: BoxPlot of the time value columns

One other thing that stood out as I analyzed all the different columns through histograms was that most of the data was not normally distributed. Fig. 5 shows the distribution of the features.

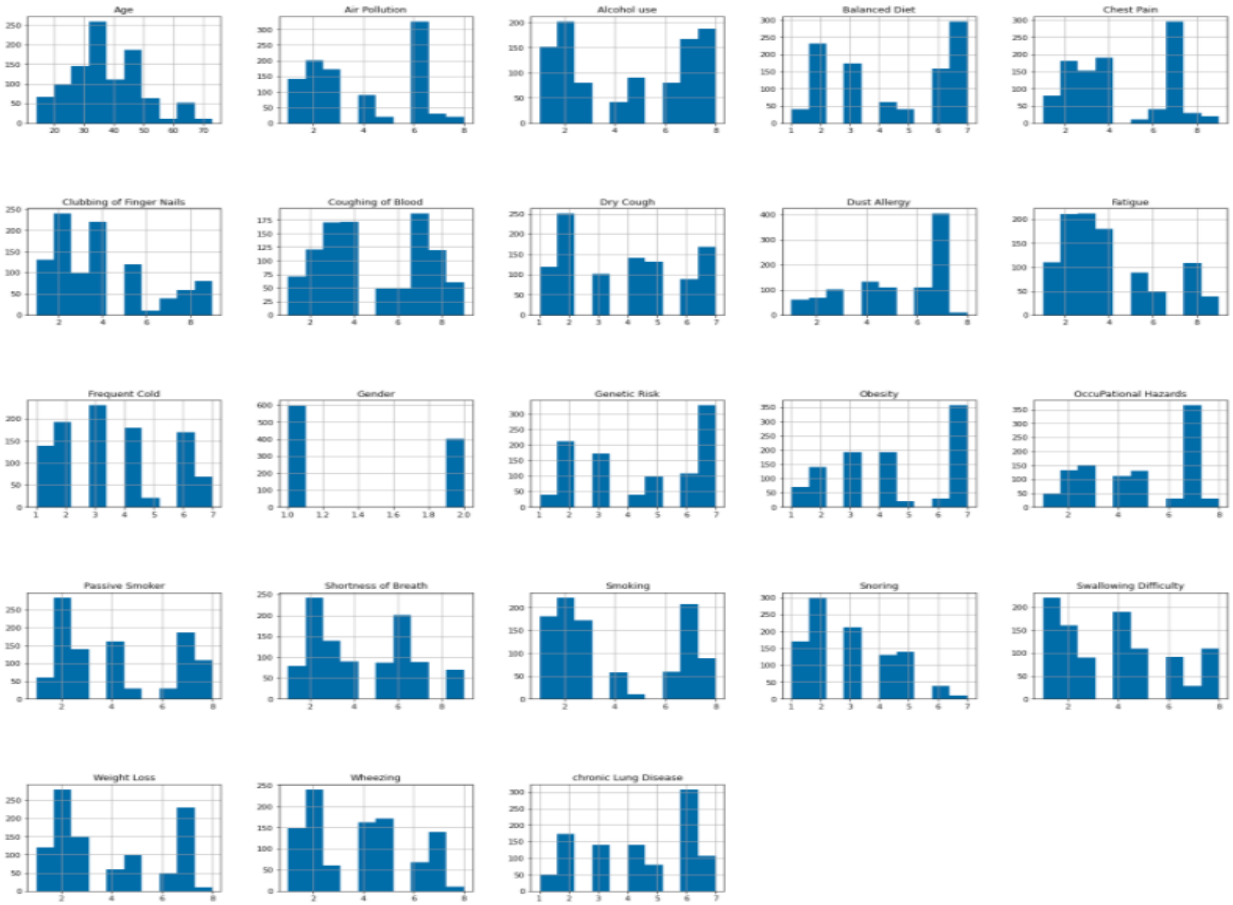


Figure 5: Histogram of distribution in different columns

Feature Engineering

Now the data is all cleaned and analysed to understand the relationships between different variables, I need to create and customize different features to prepare my dataset for modelling. After detailed analysis, I came to the conclusion that a lot of the feature engineering that was previously done was optimal for me to start modelling.

The main thing I did at this point was remove the feature Patient Id that would not be needed for the modelling part. Table 1 below is what I used for modelling.

	count	mean	std	min	25%	50%	75%	max
Age	1000.0	37.174	12.005493	14.0	27.75	36.0	45.0	73.0
Gender	1000.0	1.402	0.490547	1.0	1.00	1.0	2.0	2.0
Air Pollution	1000.0	3.840	2.030400	1.0	2.00	3.0	6.0	8.0
Alcohol use	1000.0	4.563	2.620477	1.0	2.00	5.0	7.0	8.0
Dust Allergy	1000.0	5.165	1.980833	1.0	4.00	6.0	7.0	8.0
OccuPational Hazards	1000.0	4.840	2.107805	1.0	3.00	5.0	7.0	8.0
Genetic Risk	1000.0	4.580	2.126999	1.0	2.00	5.0	7.0	7.0
chronic Lung Disease	1000.0	4.380	1.848518	1.0	3.00	4.0	6.0	7.0
Balanced Diet	1000.0	4.491	2.135528	1.0	2.00	4.0	7.0	7.0
Obesity	1000.0	4.465	2.124921	1.0	3.00	4.0	7.0	7.0
Smoking	1000.0	3.948	2.495902	1.0	2.00	3.0	7.0	8.0
Passive Smoker	1000.0	4.195	2.311778	1.0	2.00	4.0	7.0	8.0
Chest Pain	1000.0	4.438	2.280209	1.0	2.00	4.0	7.0	9.0
Coughing of Blood	1000.0	4.859	2.427965	1.0	3.00	4.0	7.0	9.0
Fatigue	1000.0	3.856	2.244616	1.0	2.00	3.0	5.0	9.0
Weight Loss	1000.0	3.855	2.206546	1.0	2.00	3.0	6.0	8.0
Shortness of Breath	1000.0	4.240	2.285087	1.0	2.00	4.0	6.0	9.0
Wheezing	1000.0	3.777	2.041921	1.0	2.00	4.0	5.0	8.0
Swallowing Difficulty	1000.0	3.746	2.270383	1.0	2.00	4.0	5.0	8.0
Clubbing of Finger Nails	1000.0	3.923	2.388048	1.0	2.00	4.0	5.0	9.0
Frequent Cold	1000.0	3.536	1.832502	1.0	2.00	3.0	5.0	7.0
Dry Cough	1000.0	3.853	2.039007	1.0	2.00	4.0	6.0	7.0
Snoring	1000.0	2.926	1.474686	1.0	2.00	3.0	4.0	7.0

Table 1: Final feature engineered dataset

Model Selection

The objective is to predict the risk of lung cancer. The different stages of risk are divided into 3 levels; Low, Medium, High. The prediction will be based on genetic risk, age, gender, lifestyle and other side effects. So for the modelling section my goal was to build the optimal model that would have the highest accuracy of predicting the Level of risk for lung cancer.

Instead of trying different individual models, I used PyCaret to run 14 different models simultaneously and rank them based on their accuracy.

Table 2 below shows the different models ranked based on their accuracy score.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.4370
nb	Naive Bayes	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0080
dt	Decision Tree Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0070
ridge	Ridge Classifier	1.0000	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0070
rf	Random Forest Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0670
qda	Quadratic Discriminant Analysis	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0090
gbc	Gradient Boosting Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.1380
et	Extra Trees Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0690
lightgbm	Light Gradient Boosting Machine	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0350
catboost	CatBoost Classifier	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.1700
svm	SVM - Linear Kernel	0.9959	0.0000	0.9958	0.9962	0.9958	0.9939	0.9940	0.0090
knn	K Neighbors Classifier	0.9795	0.9976	0.9787	0.9818	0.9794	0.9692	0.9704	0.0130
lda	Linear Discriminant Analysis	0.9243	0.9551	0.9240	0.9322	0.9243	0.8861	0.8899	0.0110
ada	Ada Boost Classifier	0.7014	0.8652	0.7012	0.8155	0.6277	0.5563	0.6478	0.0310

Table 2: All of the different models ranked based on accuracy

From the table above, I chose 3 different machine learning classification models: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier. The

metric I focused on when building my models was accuracy. I wanted my model to predict the level of risk of a patient. Luckily most of the models gave almost 100% accuracy in predicting the risk level of lung cancer.

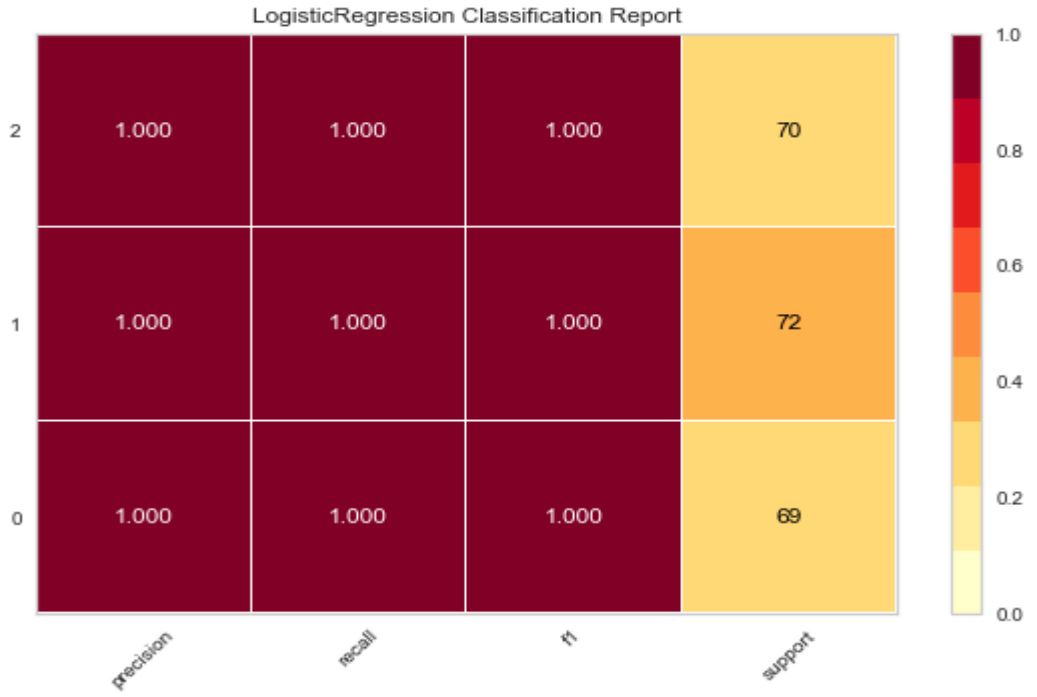


Figure 6: LogisticRegression Report

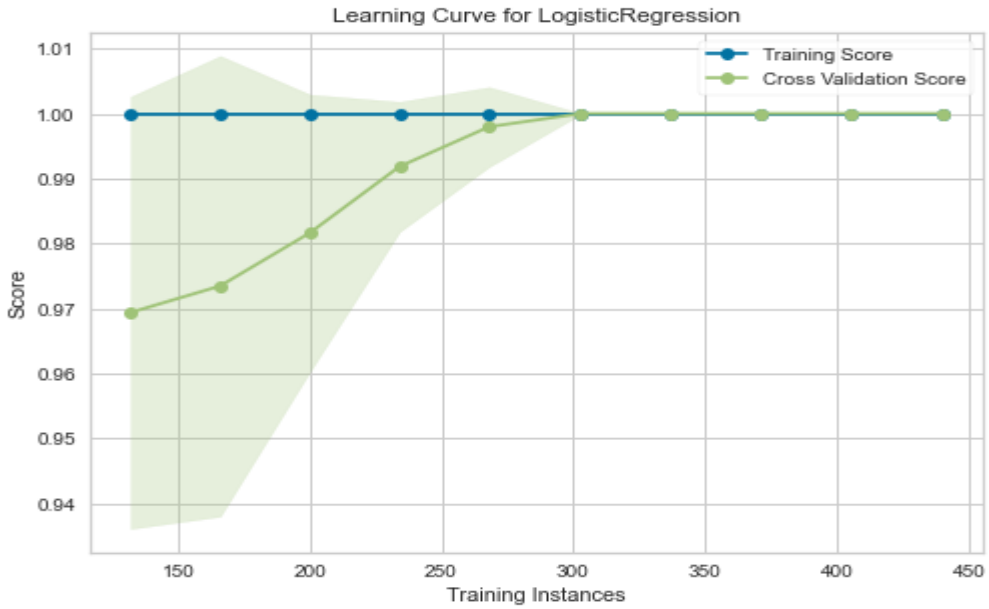


Figure 7: LogisticRegression Learning Curve

Before building the models, I would like to share the importance of different features that have the most impact on the level of risk. Figure 6 shows the top features that have the most impact. In my initial analysis, I did not put too much emphasis on how important obesity, coughing of blood and passive smoking were, however, we see that these were very important. Anything above 30, I would consider as of high importance.

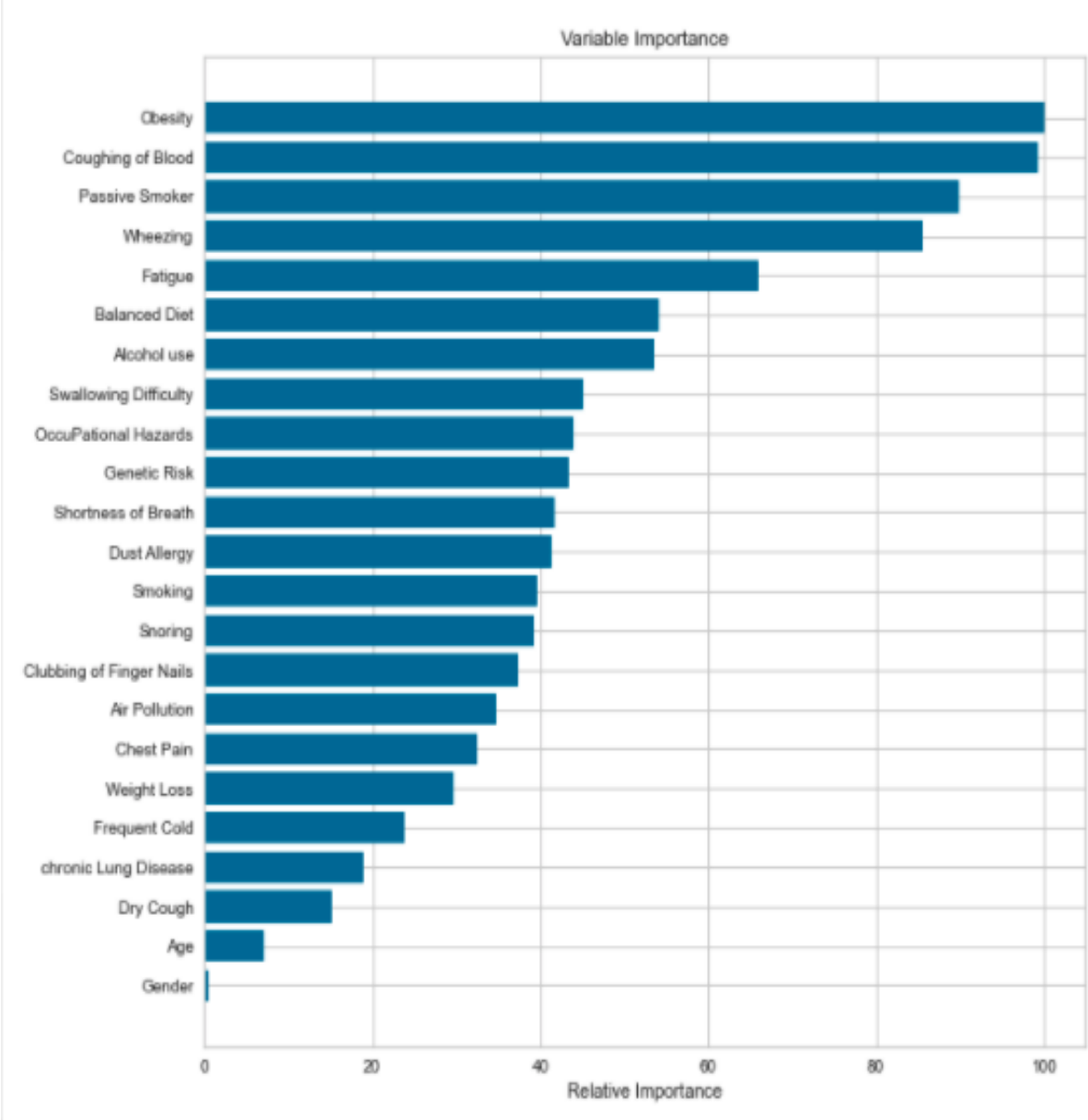


Figure 6: Feature importance

Takeaways

We see that Logistic Regression is the best model for predicting the risk of lung cancer in patients. Although the majority of the models had very good accuracy, the Logistic Regression model consistently came as number one due to the higher **TT (Sec.**

As stated in the beginning, as the cancer progresses and continues to spread, just like other forms of cancers, it becomes harder to treat and the survival rate decreases significantly. By diagnosing a patient as soons as possible we can add years if not decades to a patient's life, with the current advancement of immuno and chemo therapies. For example, a few extra months or years could be enough time for a new drug or therapy to be developed or optimized to be approved by the FDA and become accessible to patients.

Not all emphasis and hoep should be put onto doctors and drug companies. As seen with feature importance, our health care professionals and education institutes should emphasize the power of personal, life decisions to prevent the increased risk of lung cancer. We should make trials and data accessible, in an understandable way, to show the benefits of conscious lifestyle habits.