**Lung Cancer Stage prediction with machine learning**

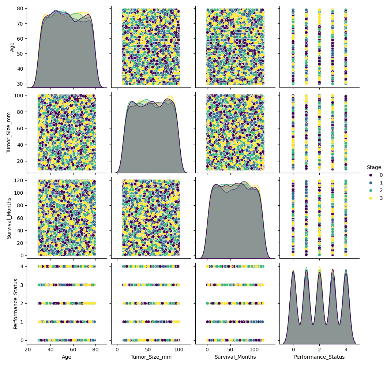
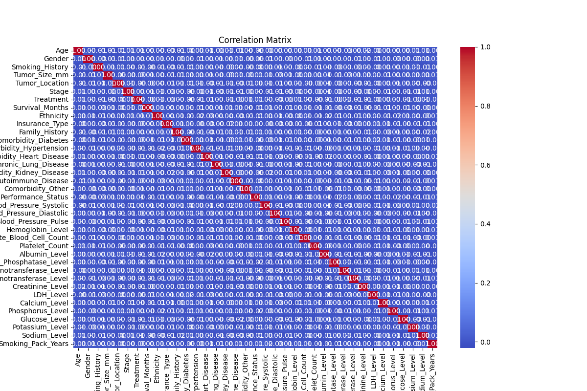
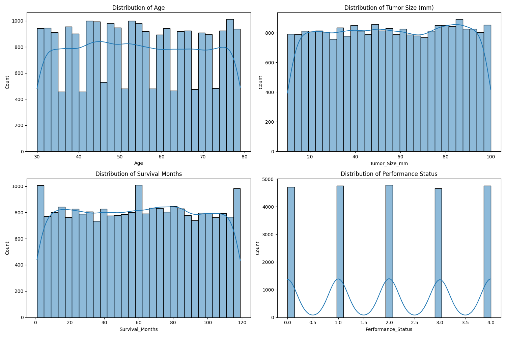
**Vladimir Gološin, Kristina Andrijin, Stefan Bogdanović**

1. Motivation

Doctors are in short supply and those who are available are often overworked and tired, especially the lung cancer specialists. In order to help them out a machine learning model can be created to aid in cancer diagnostics.

2. Research questions

Specifically, we want to make a machine learning model for classification of lung cancer stages based on various parameters, these include: Age, Gender, Smoking\_History, Tumor\_Size\_mm, Tumor\_Location, Stage, Treatment, Survival\_Months, Ethnicity, Insurance\_Type, Family\_History, Comorbidity\_Diabetes, Comorbidity\_Hypertension, Comorbidity\_Heart\_Disease, Comorbidity\_Chronic\_Lung\_Disease, Comorbidity\_Kidney\_Disease, Comorbidity\_Autoimmune\_Disease, Comorbidity\_Other, Performance\_Status, Blood\_Pressure\_Systolic, Blood\_Pressure\_Diastolic, Blood\_Pressure\_Pulse, Hemoglobin\_Level, White\_Blood\_Cell\_Count, Platelet\_Count, Albumin\_Level, Alkaline\_Phosphatase\_Level, Alanine\_Aminotransferase\_Level, Aspartate\_Aminotransferase\_Level, Creatinine\_Level, LDH\_Level, Calcium\_Level, Phosphorus\_Level, Glucose\_Level, Potassium\_Level, Sodium\_Level, Smoking\_Pack\_Years. We will now take a look at how these are distributed among the cancer stages.



We can see that the data is very equally distributed among the classes. Besides that, the data has a very low correlation with each other, it is around 0. In the final set of graphs we can see that the data has a rather uniform distribution, which means it is most likely synthetic in origin.

3. Related work

There are 8 code submissions on Kaggle for this data set. 5 of them are an analysis of the data set, and they have come to similar conclusions to us. The data set is very balanced, each of the categorical labels are exactly equal, for example all of the Stage categories take up ~25%, there is almost an exactly 50/50 split between men and women, there is 20% of each ethnicity and so on.

Other 3 code submissions are attempts at solving the problem. They involve encoding the categorical columns, dropping irrelevant columns and other preprocessing techniques. They used a variety of classifiers and composition classifiers such as AdaBoost, CatBoost, LightGBM, Perceptron, Ridge, Random Forest, Decision Tree and so on… all of them got a macro F1 score result around 0.25

4. Methodology

We have approached the problem in a multitude of ways. We will describe them separately  
Approach 1:

Approach 1 included engineering features, specifically merging Comorbidity columns into a single Comorbidity\_Count column, as well as Label Encoding other categorical columns. A number of columns were dropped since it was determined for them not to have a good impact on the final result, these were: Hemoglobim\_Level,White\_Blood\_Cell\_Count, Platelet\_Count, Aspartate\_Aminotransferase\_Level, Creatinine\_Level, LDH\_Level, Calcium\_Level, Insurance\_Type, as well as the Patient\_ID. SimpleImputer has been used to add missing data, but I’m not even sure there was any missing data. StandardScaler was used to scale the remaining columns.

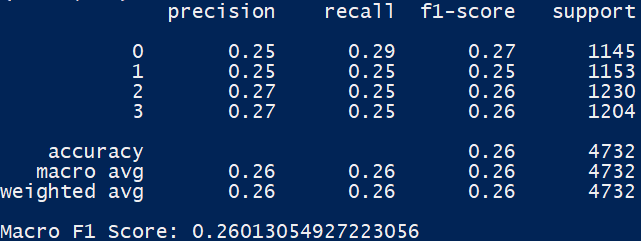
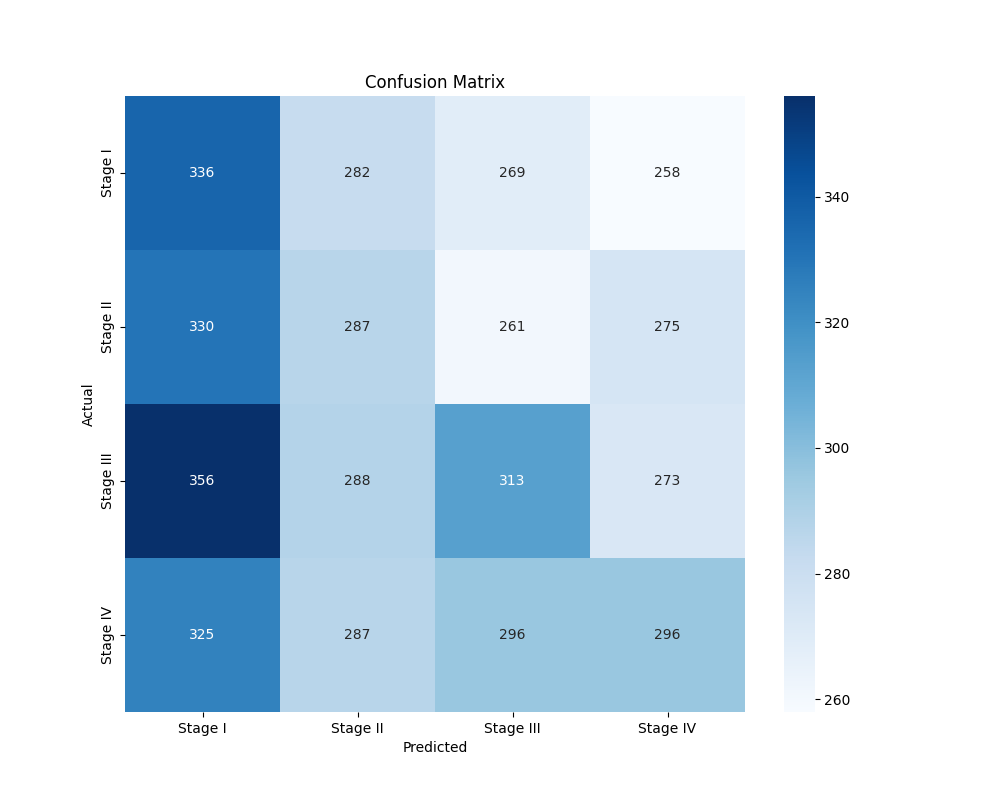
Approach 2:

A Grid Search with 5-fold cross-validation was conducted to optimize hyperparameters of Random Forest Classificators. The best parameters found were: max\_depth=20, min\_samples\_leaf=2, min\_samples\_split=10, and n\_estimators=100. The model achieved a macro F1 score of 0.2563 on the test set. Feature importance analysis and confusion matrix were generated for further insights.

5. Discussion

Again as with the last chapter, we will discuss the approaches separately.

Approach 1:  
 The data set has been split 80:20 into a train set and a test set. We have used the train set with a SVC classifier which had the following hyperparameters: kernel=’rbf’, gamma=’scale’, C=5. Rbf was used as we have found the data set to be rather non-linear. Other two were found with trial and error by hand. Final F1 Macro score was 0.2601 which is around the 0.25 others have been getting. I will blame the data set for this low score as it is very synthetic and possibly even random generated.



We can see that the model heavily favours Stage 1, and is a little bit more precise with Stage 4.

Approach 2:

Next classification attempt was performed using Decision Tree, KNN, and SVM classifiers, optimized through Grid Search. Data preprocessing included encoding categorical features and standardizing numerical features. Grid Search with 5-fold cross-validation determined optimal hyperparameters: Decision Tree (max\_depth=30, min\_samples\_leaf=1, min\_samples\_split=5), KNN (n\_neighbors=10, weights=distance), and SVM (C=5, gamma=scale, kernel=rbf). These models were combined in a VotingClassifier with soft voting. The ensemble achieved a macro F1 score of 0.2492 on the test set.

6. References

Data set: <https://www.kaggle.com/datasets/rashadrmammadov/lung-cancer-prediction/data>

Code submissions: <https://www.kaggle.com/datasets/rashadrmammadov/lung-cancer-prediction/code>