Prediction of Lung Cancer Stage

ANJA PETKOVIĆ SV22/2020, BRANISLAV STOJKOVIĆ SV64/2020

1.MOTIVATION

Since we dealt with the same topic in another subject using rule-based systems, we were motivated to explore it further through machine learning. During the development of that project, we consulted with medical experts specializing in lung cancer. From their insights, we realized that not every patient can be diagnosed using a linear set of rules. Each patient requires special consideration and analysis of their findings. Therefore, we found that machine learning could be used to create such a system.

2. RESEARCH QUESTIONS

While determining the topic and dataset for the project, we encountered various datasets. Some had fewer rows and columns, while others were more extensive. We aimed for a dataset more complex than those used in exercises, with as many rows and columns as possible. After considering these criteria, we found a suitable dataset available at the following link: Kaggle Lung Cancer Prediction Dataset.

The dataset includes the following attributes:

- **Patient_ID**: Patient identification number.
- **Age**: Age of the patient.
- **Gender**: Gender of the patient.
- **Smoking_History**: Smoking history of the patient (e.g., Current Smoker, Never Smoked, Former Smoker).
- **Tumor_Size_mm**: Tumor size in millimeters.
- **Tumor_Location**: Tumor location (e.g., Lower Lobe).
- **Stage**: Cancer stage (e.g., Stage I, Stage II, Stage III).
- **Treatment**: Treatment received by the patient (e.g., Surgery, Radiation Therapy, Chemotherapy).
- **Survival_Months**: Number of survival months after diagnosis.
- **Ethnicity**: Ethnicity of the patient.
- **Insurance_Type**: Type of insurance (e.g., Medicare, Private, Other).
- **Family_History**: Family history of cancer (Yes/No).
- **Comorbidity_Diabetes**: Presence of diabetes (Yes/No).
- **Comorbidity_Hypertension**: Presence of hypertension (Yes/No).
- **Comorbidity_Heart_Disease**: Presence of heart disease (Yes/No).
- **Comorbidity_Chronic_Lung_Disease**: Presence of chronic lung disease (Yes/No).
- **Comorbidity_Kidney_Disease**: Presence of kidney disease (Yes/No).
- **Comorbidity_Autoimmune_Disease**: Presence of autoimmune diseases (Yes/No).
- **Comorbidity_Other**: Presence of other comorbidities (Yes/No).
- **Performance_Status**: Performance status of the patient (scale from 0 to 4).

- **Blood_Pressure_Systolic**: Systolic blood pressure.
- **Blood_Pressure_Diastolic**: Diastolic blood pressure.
- **Blood_Pressure_Pulse**: Pulse.
- **Hemoglobin_Level**: Hemoglobin level.
- **White_Blood_Cell_Count**: White blood cell count.
- **Platelet_Count**: Platelet count.
- **Albumin_Level**: Albumin level.
- **Alkaline_Phosphatase_Level**: Alkaline phosphatase level.
- **Alanine_Aminotransferase_Level**: Alanine aminotransferase level.
- **Aspartate_Aminotransferase_Level:** Aspartate aminotransferase level.
- **Creatinine_Level**: Creatinine level.
- **LDH_Level**: Lactate dehydrogenase level.
- **Calcium_Level**: Calcium level.
- **Phosphorus_Level**: Phosphorus level.
- **Glucose_Level**: Glucose level.
- **Potassium_Level:** Potassium level.
- **Sodium_Level**: Sodium level.
- **Smoking_Pack_Years**: Number of smoking years (expressed in pack-years).

3. RELATED WORK

Based on this dataset, it is possible to predict several columns, such as the cancer stage, the number of months the patient will survive, and the type of therapy. Most solutions focus on predicting the cancer stage and the number of months the patient will survive. The preprocessing steps for these classifications typically include removing irrelevant features, encoding features, and various types of normalization. After data preparation, models are built. The models used include SVC, RandomForestClassifier, AdaBoostClassifier, DecisionTreeClassifier, KNeighborsClassifier, XGBClassifier, CatBoostClassifier, LGBMClassifier, StackingClassifier, RidgeClassifier, Perceptron, and ExtraTreesClassifier.

4. METHODOLOGY

To address the problem of predicting the stage of lung cancer, we followed a systematic approach involving data preprocessing, feature engineering, model training, and evaluation. Here is a detailed explanation of the methodology:

4.1 DATA PREPROCESSING

- **Data loading**: The dataset was loaded using pandas.
- **Feature selection**: We dropped irrelevant features that were not contributing to the stage prediction, such as **Survival_Months**, **Blood_Pressure_Systolic**, **Blood_Pressure_Diastolic**, **Age**, **Blood_Pressure_Pulse**, **Performance_Status, Gender, Ethnicity, Insurance_Type** and **Patient_ID**. Features were dropped based on heatmap and classification report.
- **Splitting data**: The dataset was split into training and testing sets using an 80-20 split.

4.2 Feature engineering

- **Categorical features**: Included Smoking_History, Tumor_Location, Treatment, Ethnicity, Insurance_Type, and Family_History.
- **Numerical features**: Included Tumor_Size_mm, 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, and Smoking_Pack_Years.

4.3 Data transformation pipelines

- **Numerical transformer**: Applied median imputation followed by standard scaling.
- **Categorical transformer**: Applied most frequent imputation followed by one-hot encoding.

4.4 Model training

- ***Gradient Boosting Classifier***:
  - **Training**: A Gradient Boosting Classifier was trained using the preprocessed data.
  - **Evaluation**: The model's performance was assessed using macro and micro F1 scores, precision and recall.
- ***Stacking Classifier***:
  - **Training**: A stacking classifier combining SVM, KNN, and Gaussian Naive Bayes with Logistic Regression as the final estimator was trained.
  - **Evaluation**: The model's performance was assessed using macro and micro F1 scores, precision and recall.

4.5 Model evaluation

- ***Evaluation metrics***: macro and micro F1 score, precision, recall.
- ***Hyperparameter optimization***: performed using Random Search algorithm.

5. Discussion

For testing the results, we used the test dataset, which is 20% of the initial dataset. As performance metrics, we used the Micro and Macro F1 scores, as well as Precision and Recall metrics. Parameter optimization was carried out using the RandomSearch algorithm. Table 1 shows the metric values for both classifiers. Picture 1 represents the classification report for Gradient Boosting classifier. Picture 2 represents the classification report for Stacking classifier.

Table 1 Results

CLASSIFIER	MICRO F1	MACRO F1	PRECISION	RECALL
BOOSTING	0.254	0.250	0.257	0.256
STACKING	0.251	0.207	0.44	0.256

Boosting - Classification Report:

	precision	recall	f1-score	support
Stage I	0.25	0.23	0.24	1170
Stage II	0.28	0.16	0.21	1233
Stage III	0.27	0.26	0.26	1175
Stage IV	0.24	0.37	0.29	1154
accuracy			0.25	4732
macro avg	0.26	0.26	0.25	4732
weighted avg	0.26	0.25	0.25	4732

Picture 1 Classification report for boosting model

Stacking - Classification Report:

	precision	recall	f1-score	support
Stage I	0.27	0.20	0.23	1170
Stage II	1.00	0.00	0.00	1233
Stage III	0.25	0.27	0.26	1175
Stage IV	0.24	0.56	0.34	1154
accuracy			0.25	4732
macro avg	0.44	0.26	0.21	4732
weighted avg	0.45	0.25	0.20	4732

Picture 2 Classification report for stacking model

KEY OBSERVATIONS FROM THE RESULTS:

- ***Gradient Boosting Classifier***: This model demonstrated relatively balanced performance across different stages, with no single stage showing extreme values in terms of precision or recall. This suggests that the model can consistently identify instances of each stage to some extent.
- ***Stacking Classifier***: This model showed a high precision but very low recall for Stage II, meaning it predicted Stage II very accurately when it did predict it, but it rarely predicted Stage II. This imbalance indicates that while the model has the potential for accurate predictions, it may require further tuning or balancing to improve recall and overall performance.

6. REFERENCES

HTTPS://WWW.KAGGLE.COM/DATASETS/RASHADRMAMMADOV/LUNG-CANCER-PREDICTION