Introduction: Overview of the Data, Predictive Task, and Summary Findings

Lung cancer remains one of the most challenging and lethal diseases worldwide, demanding accurate predictive models to assist medical professionals in decision-making. For this project, we leveraged a publicly available dataset from the University of California, Irvine (UCI) Machine Learning Repository, which contains patient data collected at the Wroclaw Thoracic Surgery Center from 2007 to 2011. This dataset is part of the National Lung Cancer Registry in Warsaw, Poland, and provides a detailed record of patients who underwent major lung resections for primary lung cancer.

The dataset consists of 16 features, capturing both clinical indicators and pre-surgery conditions, including:

- Patient demographics (e.g., Age),
- Lung capacity metrics (PRE4 for forced vital capacity and PRE5 for the volume exhaled in the first second of forced expiration),
- Symptoms and pre-existing conditions (e.g., coughing, physical weakness, asthma)
- Tumor-specific attributes (e.g., size and type).

An initial exploration of the data revealed a strong class imbalance, with only 14.89% of patients not surviving beyond the first year. This imbalance presents a challenge for model accuracy, particularly for identifying at-risk patients.To address this, several machine learning models were applied, including:

- Random Forest Classifier
- XGBoost Classifier
- Support Vector Machine (SVC)
- Ridge Classifier
- LightGBM + SMOTE
- Baseline Model for comparison

Among these, XGBoost and Random Forest emerged as the most promising, with XGBoost demonstrating better recall for identifying patients at higher risk of mortality, although with some trade-offs in overall accuracy.

The goal of this project is to provide a data-driven predictive tool that can enhance clinical decision-making by identifying high-risk patients more accurately, potentially guiding more personalized treatment plans and improving post-surgical care strategies.

Data Description: Data Source and Description

The dataset used for this project originates from the University of California, Irvine (UCI) Machine Learning Repository, specifically from the Thoracic Surgery Data collection. The data was collected at the Wroclaw Thoracic Surgery Center in Poland, covering patient records from 2007 to 2011. The dataset is also a part of the National Lung Cancer Registry in Warsaw, which consolidates clinical data of lung cancer patients undergoing major lung resections.

Dataset Overview

The dataset comprises 470 instances of lung cancer patients, each represented by 16 features and a binary outcome variable (Risk1Yr) that indicates the patient's survival status one year post-surgery:

1 → Patient passed away within one year

0 → Patient survived beyond one year

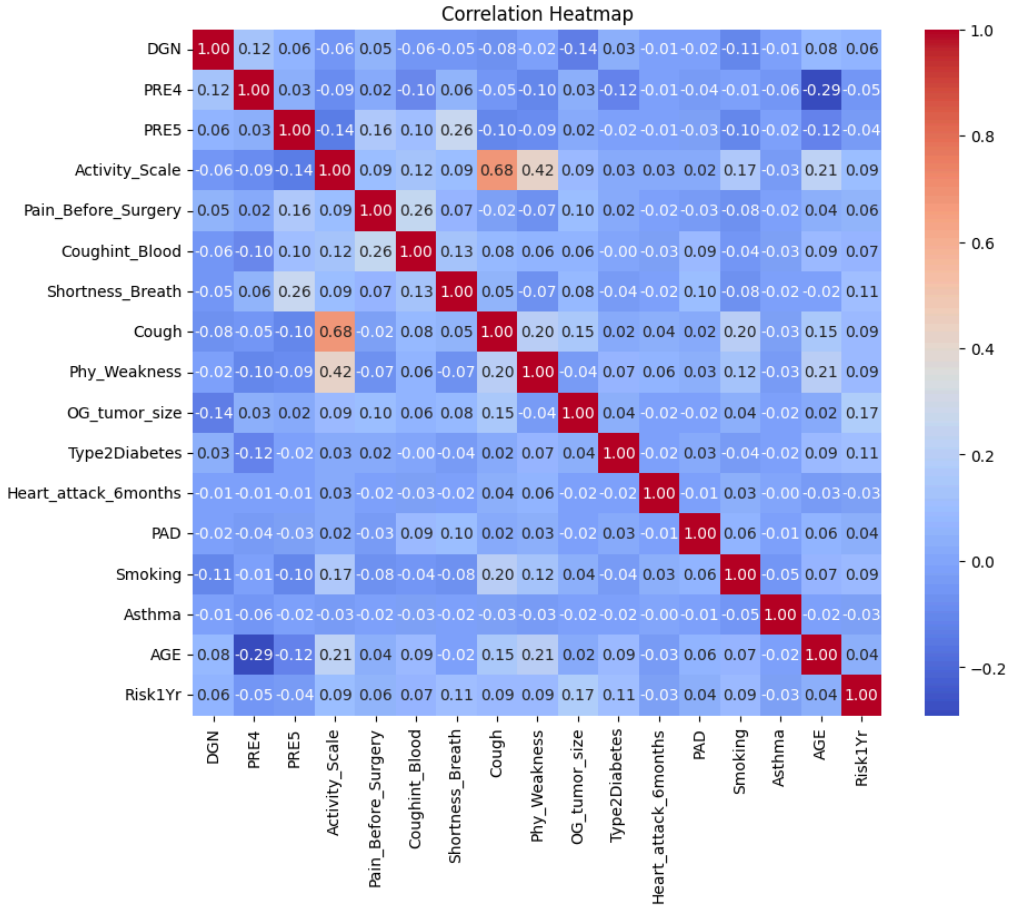
Features and Their Descriptions

The dataset contains a mix of:

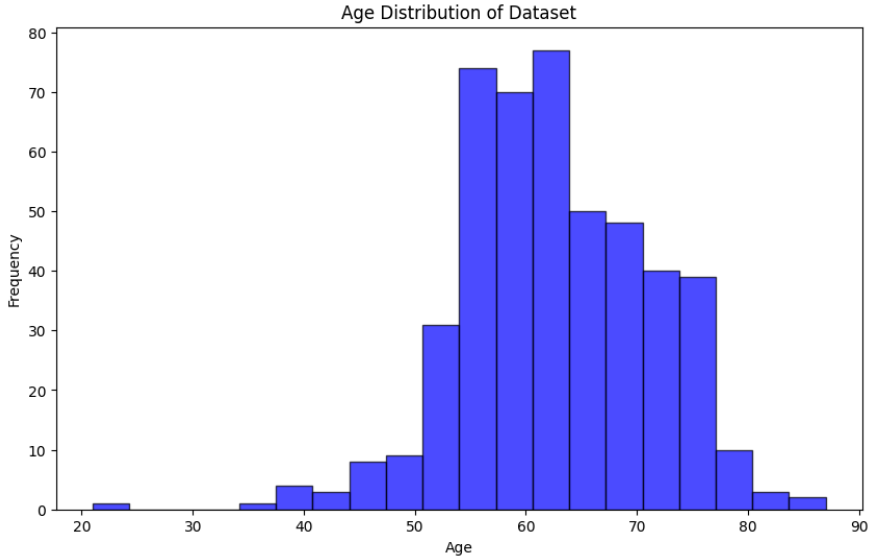
- Categorical Features: Encoded as numerical values representing different diagnostic types and physical conditions.
- Numerical Features: Continuous measurements of lung function and patient age.
- Binary Features: Indicators of specific symptoms and conditions (e.g., coughing, asthma, smoking history).

Feature Name	Type	Description
DGN	Categorical	Type and severity of the tumor, encoded from 1 to 8
PRE4	Numerical	Forced Vital Capacity (FVC) of the lungs
PRE5	Numerical	Volume exhaled in the first second of forced expiration
PRE6 to PRE32	Numerical/Binary	Various clinical measurements and symptoms
Pain Before Surgery	Binary (0/1)	Presence of physical pain prior to surgery
Coughing	Binary (0/1)	Whether the patient coughed regularly
Shortness of Breath	Binary (0/1)	Regular experience of breath shortness
Physical Weakness	Binary (0/1)	Physical weakness observed before surgery
Smoking	Binary (0/1)	Whether the patient smoked regularly
Asthma	Binary (0/1)	Diagnosis of asthma
Original Tumor Size	Categorical	The size of the tumor targeted during surgery
Type 2 Diabetes	Binary (0/1)	Whether the patient had Type 2 Diabetes
MI	Binary (0/1)	Heart attack occurrence within the last 6 months
PAD	Binary (0/1)	Diagnosis of Peripheral Arterial Disease
Haemoptysis	Binary (0/1)	Coughing up blood prior to surgery
Age	Numerical	Patient's age in years
Risk1Yr	Binary (0/1)	Survival status one year post-surgery

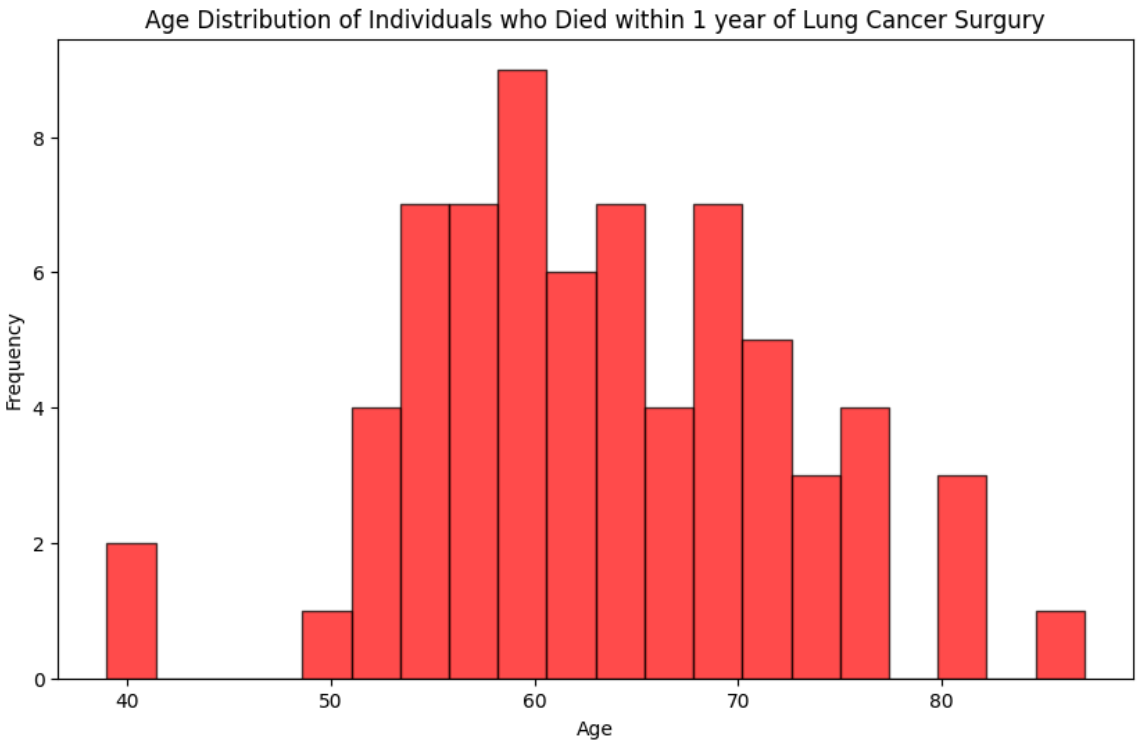
The correlation heatmap provides a visual representation of the relationships between the features in the **Thoracic Surgery Dataset**, including the target variable `Risk1Yr`. From the analysis, it is evident that certain features exhibit stronger correlations with post-surgery survival. For instance, `OG_tumor_size` shows a noticeable positive correlation (0.17) with `Risk1Yr`, suggesting that larger tumors are slightly linked with higher mortality rates. Similarly, `Activity_Scale` (0.09) and `Physical Weakness` (0.09) also demonstrate some association with increased risk, indicating that patients with limited physical activity or weakness before surgery tend to have worse outcomes. In contrast, features such as `PRE4` and `PRE5`, which represent lung capacity and volume, exhibit slight negative correlations with mortality, aligning with the clinical expectation that stronger lung function may contribute to higher survival rates. Interestingly, conditions like `PAD`, `Asthma`, and `Type2Diabetes` show very minimal correlation with `Risk1Yr`, which corresponds to the findings in the write-up where it was noted that none of the patients who passed away had asthma, reflecting its limited predictive power in this context. Additionally, strong internal correlations were observed between `Cough` and `Activity_Scale` (0.68), indicating that patients who cough frequently also tend to have reduced physical abilities. These insights suggest that certain features may hold more predictive power than others, guiding the selection of significant variables for model optimization.



The histogram displayed represents the Age Distribution of Patients in the Thoracic Surgery Dataset. From the visualization, it is evident that the majority of patients who underwent surgery were between 50 and 70 years old, with the peak frequency occurring around the 60 to 65 age range. This suggests that lung cancer surgeries are predominantly performed on middle-aged to elderly patients, which aligns with medical understanding that lung cancer risks increase significantly with age. The distribution is approximately bell-shaped, indicating a concentration of cases around the mean age, with fewer cases at the extreme age ranges (less than 40 and greater than 80). This distribution is consistent with the dataset's demographic characteristics as described in the write-up, where the average age for those who survived was around 62, and for those who did not survive, it was slightly higher at 63. This slight age difference could imply that age, although not the sole predictor, has some impact on post-surgery survival.

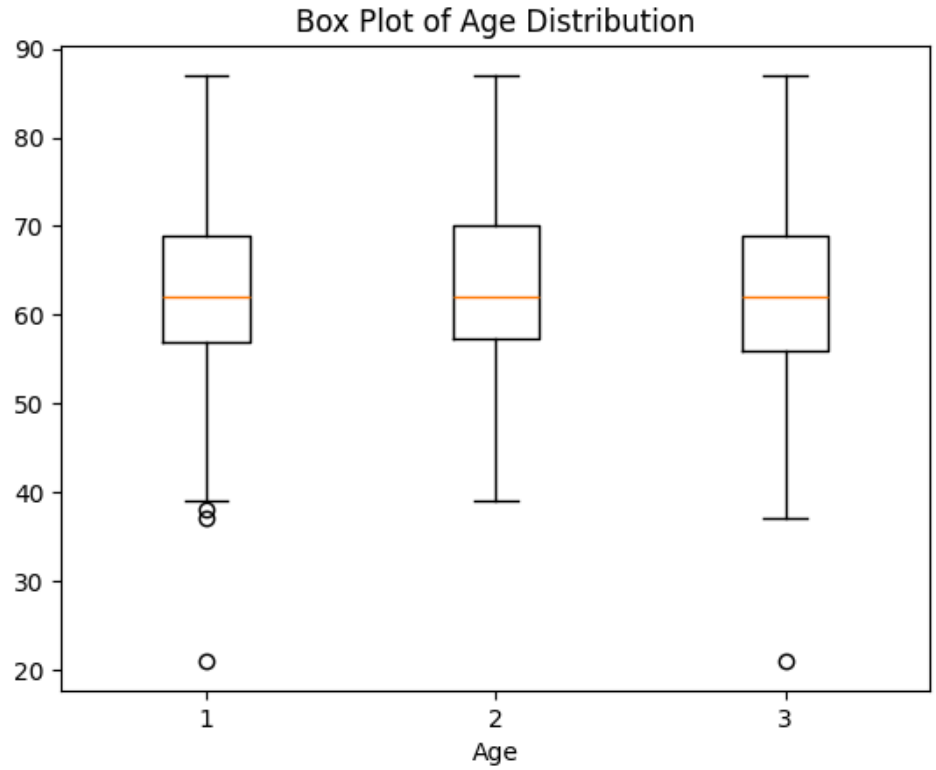


From the visualization, the highest frequency of deaths occurs primarily around the 60 to 65 age range, which is consistent with the general age distribution of lung cancer patients. This indicates that even though the majority of patients are within this age bracket, it also carries a higher risk of mortality within one year post-surgery.Interestingly, there are fewer deaths among younger patients (below 50), which may suggest that age is indeed a contributing factor to survival chances. At the older age spectrum (above 75), there are still observable cases, but their frequency is lower, possibly due to fewer elderly patients undergoing surgery in the first place.Compared to the overall age distribution you provided earlier, this chart shows a slight shift towards older ages for those who did not survive, which aligns with your write-up's statement that the average age of those who did not survive is slightly higher (63) than those who did (62).

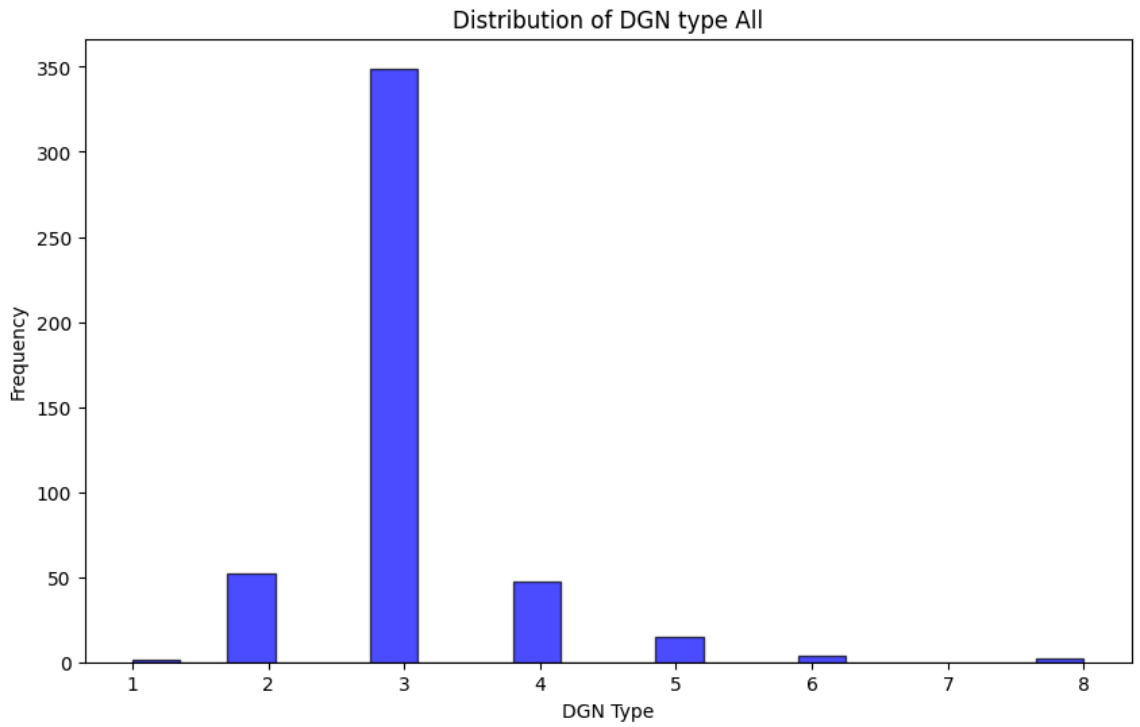


The box plot represents the age distribution across three distinct groups labeled as 1, 2, and 3, potentially corresponding to different diagnostic categories or risk levels. The median age for all three groups is consistent, around 60 years, which aligns with the average age findings described in your write-up, where the typical age for lung cancer surgery patients was noted to be approximately 62. The interquartile range (IQR), represented by the height of each box, shows a similar age spread across the groups, indicating a stable distribution of ages within each category. Notably, Group 1 has visible outliers, including a particularly young patient around 20 years old and another in the 40s, while Group 3 also presents a minor outlier. The whiskers extend to about 90 years across all groups, reflecting the presence of elderly patients in each category. These observations suggest that while age distribution remains generally consistent across categories, there are occasional exceptions, which could be linked to unique

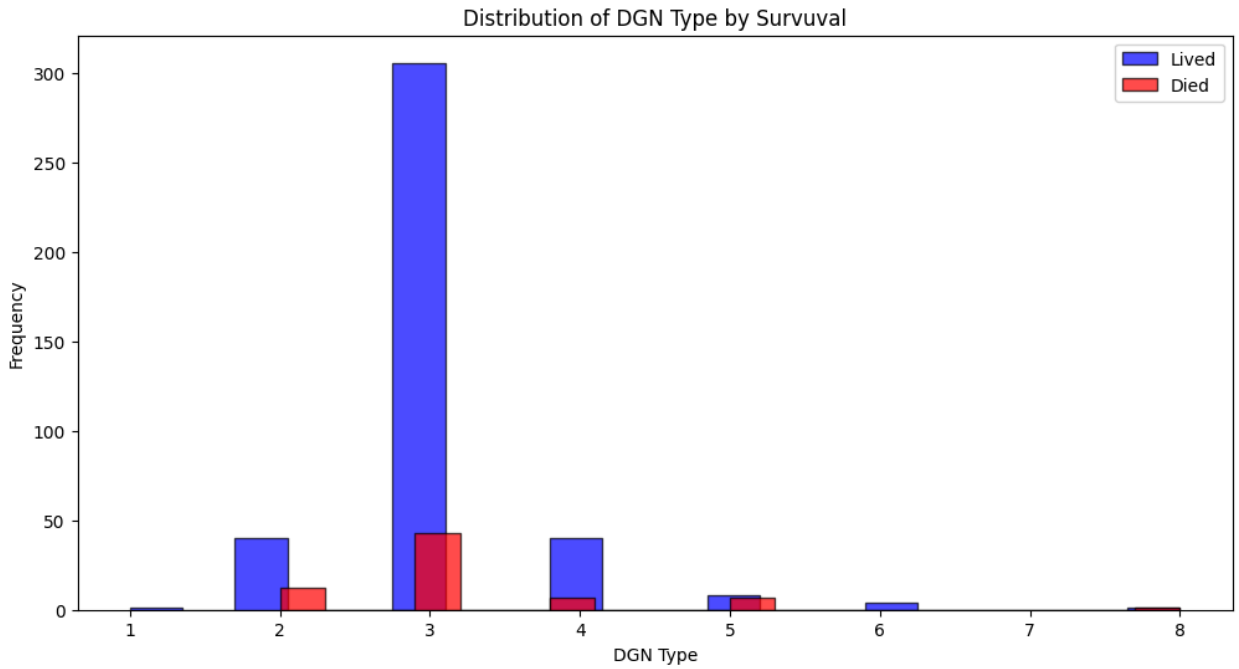
risk factors or conditions not typical of the broader group.



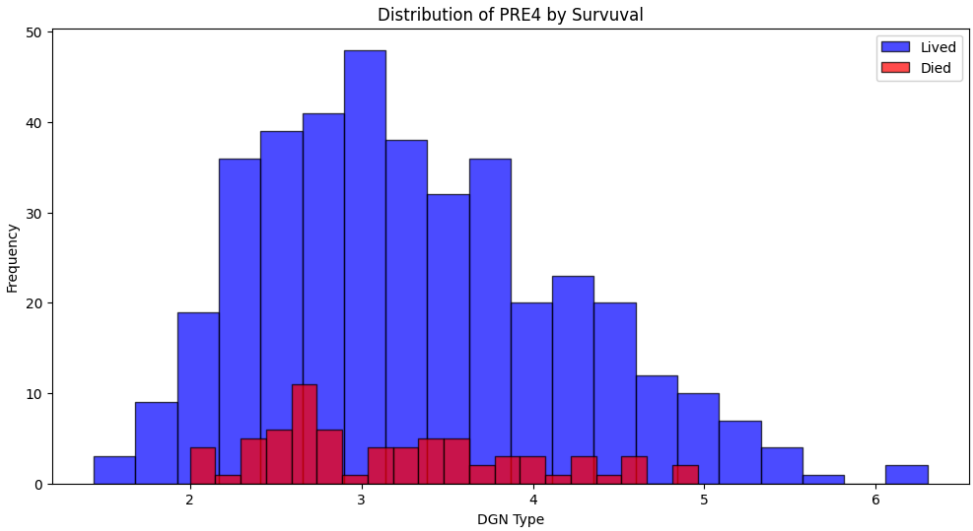
The histogram illustrates the distribution of **DGN types** across the entire dataset, representing different diagnostic categories for lung cancer severity. It is evident that the vast majority of patients belong to **DGN Type 3**, with a frequency exceeding 350 instances, while the other types are significantly less represented. Specifically, **DGN Type 2** and **DGN Type 4** are the next most common, though they account for far fewer cases compared to Type 3. This heavy concentration suggests that most lung cancer surgeries in the dataset were performed on patients with a specific tumor diagnosis classified under Type 3. In contrast, categories like **DGN Type 1, 5, 6, 7, and 8** are nearly negligible, indicating that these diagnoses are either rare or not commonly subjected to surgical intervention. This imbalance in diagnostic categories may influence model predictions, as the classifiers may become biased towards recognizing patterns associated with the most frequent types, potentially overlooking the nuances of less common diagnoses.



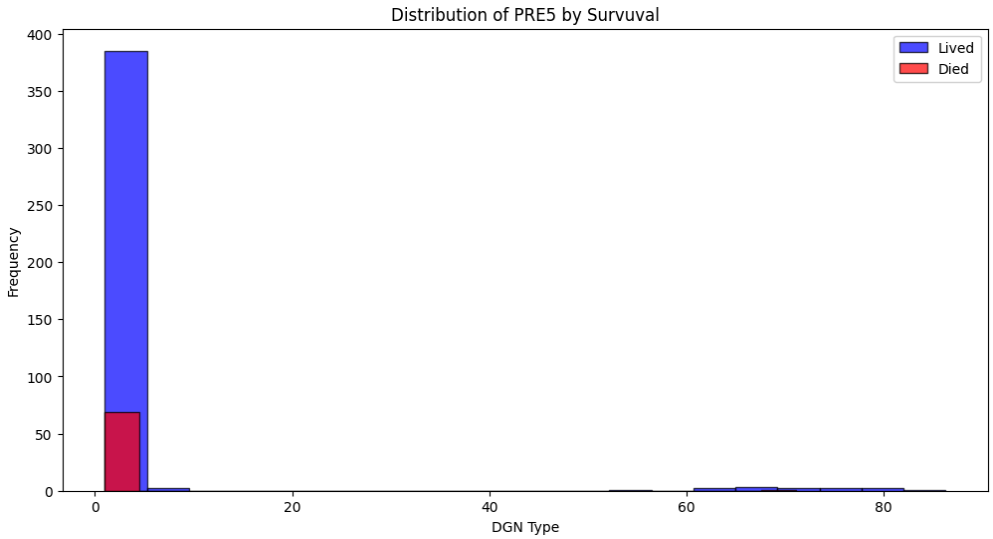
The bar plot illustrates the distribution of DGN Types (diagnostic categories) segmented by Survival Status (Lived vs. Died). The blue bars represent patients who survived beyond one year after surgery, while the red bars represent those who did not survive. It is evident that **DGN Type 3** is overwhelmingly the most common diagnosis, with the majority of patients surviving post-surgery. However, there is still a noticeable portion of patients within Type 3 who did not survive, indicating that although it is the most frequent, it still carries significant risk. For DGN Type 2 and DGN Type 4, the proportion of non-survivors (red bars) compared to survivors (blue bars) is higher relative to Type 3, suggesting these types might be linked with slightly worse prognoses. On the other hand, DGN Types 5, 6, 7, and 8 are minimally represented, with only a handful of cases recorded. The low frequency in these categories makes it challenging to draw solid conclusions about their survival outcomes, though the balance of red and blue bars hints at a less favorable prognosis.Overall, the distribution reveals that while DGN Type 3 is the most common and generally has better survival rates, specific other types may carry higher risks, potentially influencing the model's ability to accurately predict mortality based on diagnosis type.



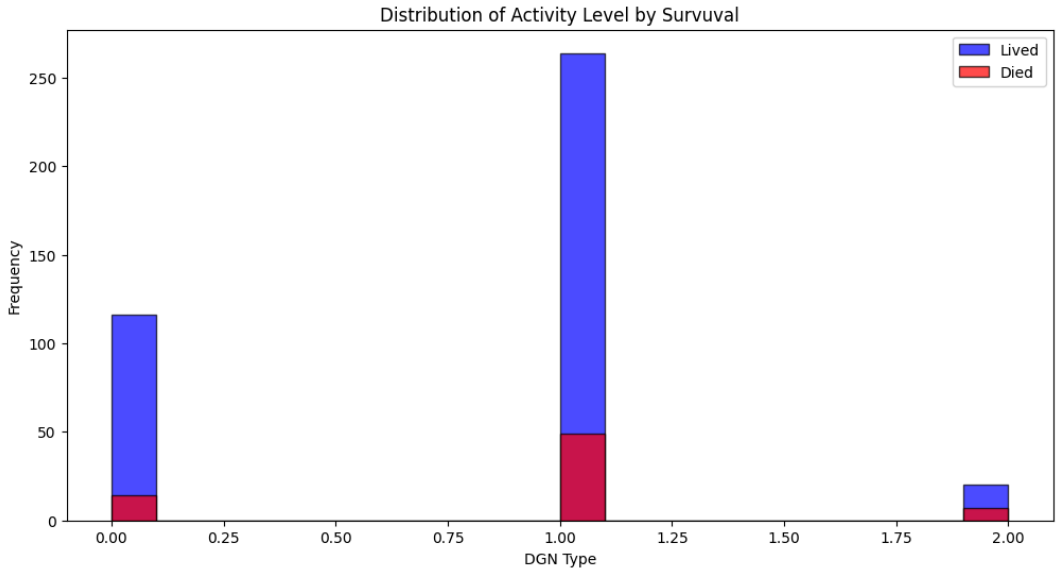
The histogram displays the distribution of PRE4 (Forced Vital Capacity)by survival status, with blue representing survivors and red representing those who did not survive. The majority of patients who survived have higher PRE4 values, generally concentrated between 2.5 and 4.5, suggesting better lung function correlates with higher survival rates. In contrast, non-survivors are more distributed across lower PRE4 values, indicating that weaker lung capacity before surgery may contribute to worse outcomes.



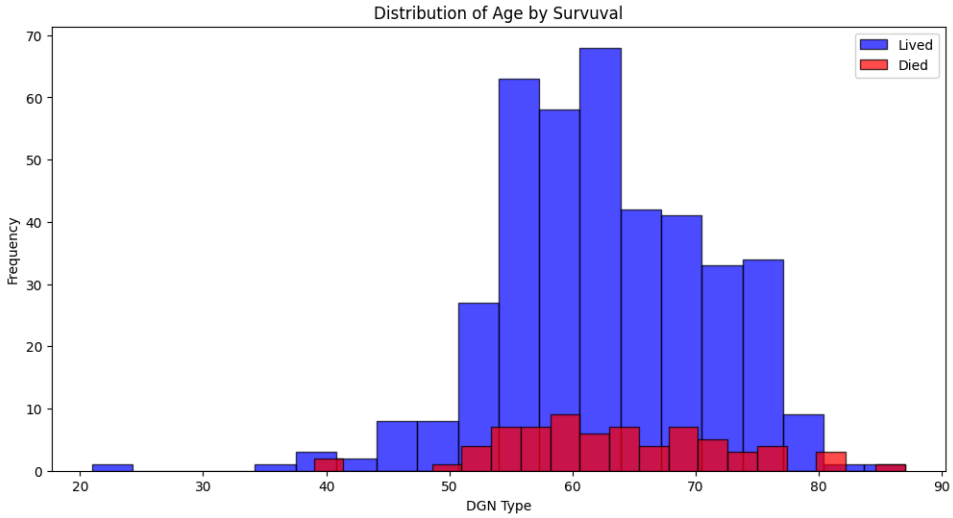
The histogram shows that most patients have **PRE5 values near 0**, with survivors (blue) outnumbering non-survivors (red). Few patients have extreme values above 20, mostly surviving.



The histogram shows the distribution of Activity Level by survival status, with 0.00, 1.00, and 2.00 representing different activity levels. Most survivors (blue) and non-survivors (red) are concentrated at 1.00 (moderate activity), while those with 0.00 (no activity) have a smaller but visible portion of non-survivors. Very few patients are at 2.00, indicating high activity is rare in the dataset.



The histogram illustrates the **Age Distribution by Survival Status**, with survivors in blue and non-survivors in red. Most patients are concentrated between **50 and 70 years old**, with survival rates generally higher in this range. However, the proportion of non-survivors increases slightly around **60 to 70 years**, reflecting the write-up's observation that the average age of those who did not survive is slightly higher. Very few patients under **40** or over **80** underwent surgery, aligning with typical age patterns for lung cancer surgeries.



Models and Methods: Overview of Models and Implementation

To address the challenge of predicting lung cancer patient survival one year after surgery, several machine learning models were employed to explore their effectiveness. The primary models used include:

- Baseline Model: This model simply predicts survival for all patients, mirroring the majority class, resulting in an accuracy of 85.11% and an F1 score of 78.26%. This benchmark was established to evaluate whether more complex models could surpass naive guessing.

- Random Forest Classifier: This model performed better than the baseline, achieving an accuracy of 86.17% and an F1 score of 80.69%. However, its ability to identify non-survivors remained limited, with a low recall of 0.07 and an F1 score of 0.13 for the minority class.

- Random Forest with Grid Search: Hyperparameter optimization was applied to the Random Forest model to enhance its predictive power. Surprisingly, the performance did not improve; the F1 score dropped back to 78.26%, matching the baseline model. Its recall for non-survivors also fell to 0, indicating a complete failure in identifying high-risk patients.

- Ridge Classifier: This linear model yielded the same performance as the Grid-Search-tuned Random Forest, with an F1 score of 78.26% and 0% recall for identifying non-survivors.

- Support Vector Classifier (SVC): Similarly to Ridge, SVC performed identically to the baseline and grid-searched Random Forest, missing all non-survivors and failing to surpass the baseline's accuracy and F1 score.

- XGBoost Classifier: Unlike the previous models, XGBoost managed to improve recall for the minority class, achieving an F1 score of 15% for predicting non-survivors and a recall of 14%. However, its overall F1 score dropped slightly to 75.84%, and it did not perform as well in identifying survivors compared to the Random Forest.

- LightGBM with SMOTE: To handle the data imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied to generate synthetic samples for the minority class. This adjustment allowed LightGBM to identify more non-survivors compared to traditional models, addressing the limitations seen in recall.

Each model was evaluated based on F1 Score to balance precision and recall, as well as Log Loss to measure the confidence of predictions. While Random Forest achieved the highest overall F1 score, XGBoost outperformed all models in Log Loss (0.67), indicating better alignment with true outcomes for non-survivors.

Results and Interpretation: Review of Modeling Results and Interpretation of Performance

The evaluation results reveal that Random Forest and XGBoost were the strongest performers among all models. Random Forest achieved the highest overall F1 score of 80.69% and showed good accuracy, while XGBoost excelled in identifying non-survivors with the lowest Log Loss of 0.67. This suggests that XGBoost, despite having a slightly lower overall F1 score, is more effective at capturing high-risk patients, making it a valuable model for predicting post-surgery mortality.

In contrast, Random Forest with Grid Search, Ridge Classifier, and SVC all performed similarly to the baseline model, failing to identify any non-survivors, which is a critical shortcoming for medical applications. The complete inability of these models to detect the minority class highlights the challenge of class imbalance in the dataset.

The Baseline Model served as a comparison point with 85.11% accuracy, but it lacked any predictive power for non-survivors. Despite this, more sophisticated models like Ridge and SVC did not outperform it, underscoring the need for more balanced data or enhanced sampling techniques.

Overall, the combination of Random Forest's stability and XGBoost's superior minority class detection positions them as the most effective models for this task. The integration of LightGBM with SMOTE also demonstrated potential in addressing imbalance, suggesting that future iterations could benefit from hybrid approaches or more advanced sampling techniques to enhance recall for non-survivors.

Conclusion and Next Steps: Summary of Findings and Future Improvements

This study aimed to predict lung cancer patient survival one year after surgery using various machine learning models. Random Forest achieved the highest overall F1 score, while XGBoost excelled in identifying non-survivors with the best Log Loss of 0.67, aligning with the

analysis of key features like PRE5 and age. However, models such as Random Forest with Grid Search, Ridge Classifier, and SVC struggled with class imbalance, failing to identify high-risk patients effectively. To improve performance, strategies like data augmentation for non-survivors, feature expansion (e.g., COPD, smoking frequency), and model ensembling are recommended. Further exploration with advanced hyperparameter tuning and deep learning architectures could enhance prediction accuracy. These improvements could make survival predictions more reliable, supporting better medical decision-making.