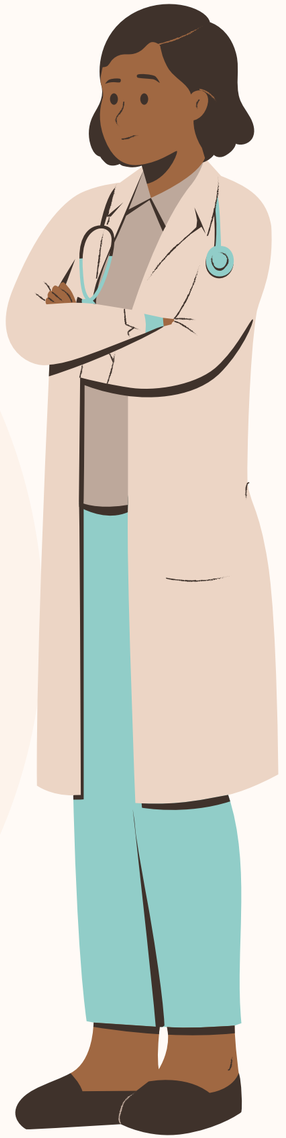
Mody University

Lung Cancer Survival Analysis Using IBM SPSS Modeler

Team Members

- **Anjali Sharma 220239**
- **Diksha Khangarot 220244**

Kiran 220247



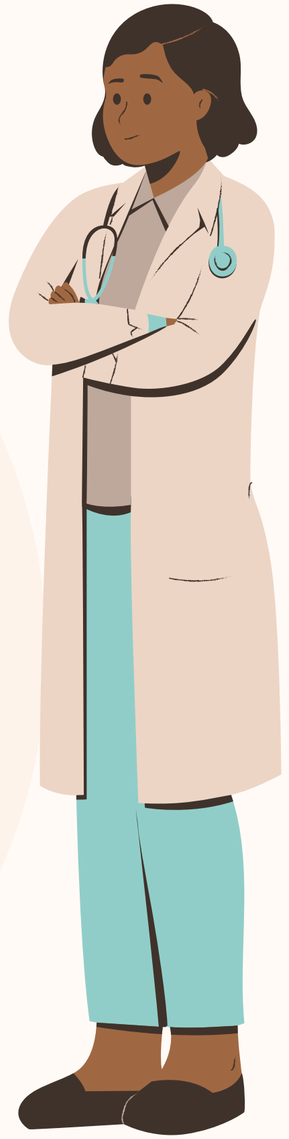
Introduction to Survival Analysis

An exploration of survival analysises for cancer

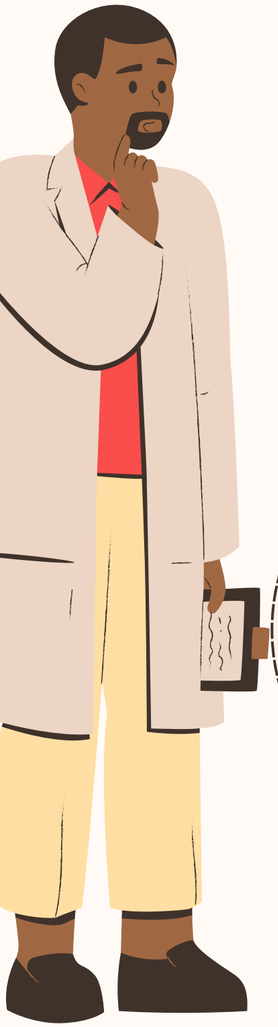
Survival analysis is a statistical method used to estimate the time until an event occurs, like death or relapse, especially in medical research. In this project, we apply survival analysis to lung cancer, one of the deadliest forms of cancer, to understand which factors influence patient survival. The goal is to use IBM SPSS Modeler to analyze a dataset of lung cancer patients and identify key variables such as treatment type, tumor type, and patient condition that impact survival outcomes. This can help in making better clinical decisions and improving patient care.

Business Understanding

Lung cancer remains one of the leading causes of cancer-related deaths globally. Early detection plays a critical role in improving survival rates and treatment outcomes. The objective of our project is to develop a low-cost and effective lung cancer prediction model using patient symptom data. This model aims to classify patients into risk categories—YES (at risk) or NO (not at risk)—based on symptoms and demographic information. The ultimate goal is to aid in early diagnosis and guide individuals toward informed healthcare decisions while understanding the key contributing factors



Data understanding



Dataset Overview

The dataset used for this project comprises 284 patient records. It includes 16 attributes: 15 predictor variables such as Age, Gender, Smoking, Yellow Fingers, and Chest Pain, and 1 target variable—Lung Cancer (YES/NO).

Data Source

The dataset was obtained from an online lung cancer prediction system, sourced from publicly available medical data repositories. This ensures reliability and relevance to real-world patient data.

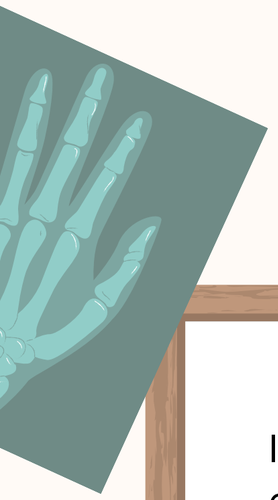
Import & Checks

We imported the dataset into IBM SPSS using .csv or .xlsx format. Preliminary checks were conducted to identify and manage any missing values or invalid entries to ensure data quality.

Visual Analysis

Data distributions were visualized using pie charts (e.g., gender ratio), bar charts (e.g., Smoking vs Lung Cancer), and histograms (e.g., Age distribution). These visualizations helped in understanding the spread and trends in the data.

Data Preparation in IBM SPSS



In IBM SPSS, we began preprocessing by encoding categorical values into numerical form. Gender was encoded as M=1 and F=2, and binary variables such as Yes/No responses were converted to Yes=2 and No=1. We addressed missing values using suitable imputation methods or, when necessary, removed incomplete entries to maintain data consistency.

Using the Variable View in SPSS, we recoded and labeled all variables appropriately, ensuring uniformity in the dataset. After cleaning and transforming the data, we finalized the independent variables—like Smoking, Yellow Fingers, Anxiety, and Chest Pain—and set Lung Cancer as the target variable. This prepared dataset was then ready for model building.

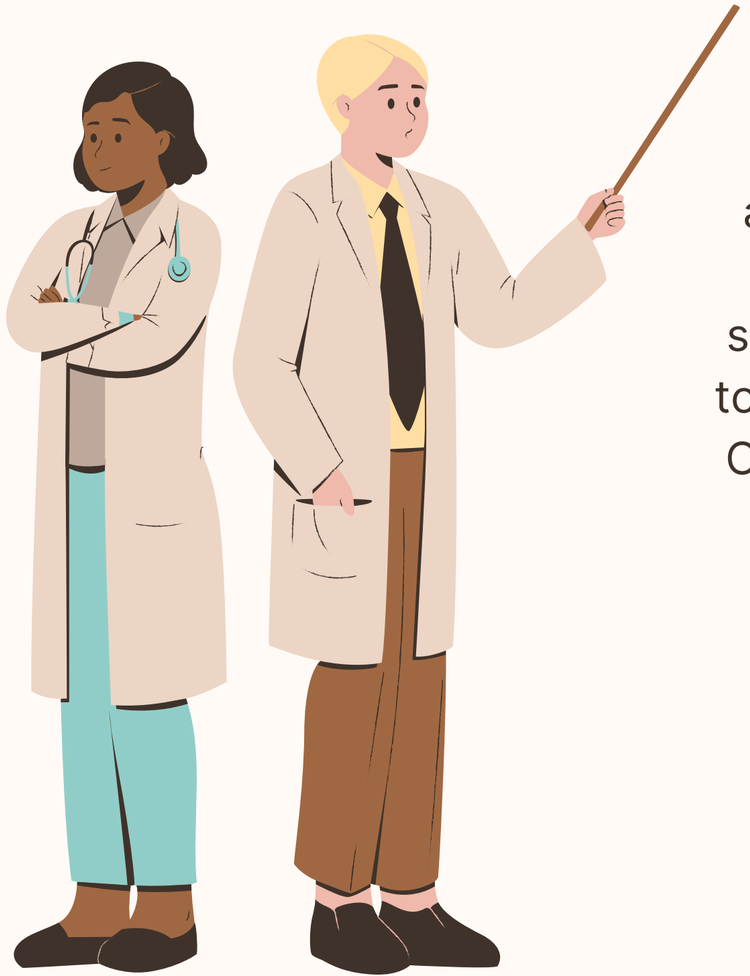
Data Modeling

Model Selection

We used Random Forest in IBM SPSS as our target variable (Lung Cancer) is binary (YES/NO). This model helps in estimating the probability of lung cancer based on input symptoms and demographics.

Modeling Steps

Using *Analyze → Random forest* we set Lung Cancer as the dependent variable and selected symptoms like Smoking, Chest Pain, and Anxiety as independent variables.



Evaluation Matrics

The model was evaluated using accuracy, Hosmer-Lemeshow test (fit), Wald test (predictor significance), and ROC Curve (AUC) to assess classification performance. Optionally, a Decision Tree was also tested for visual understanding.

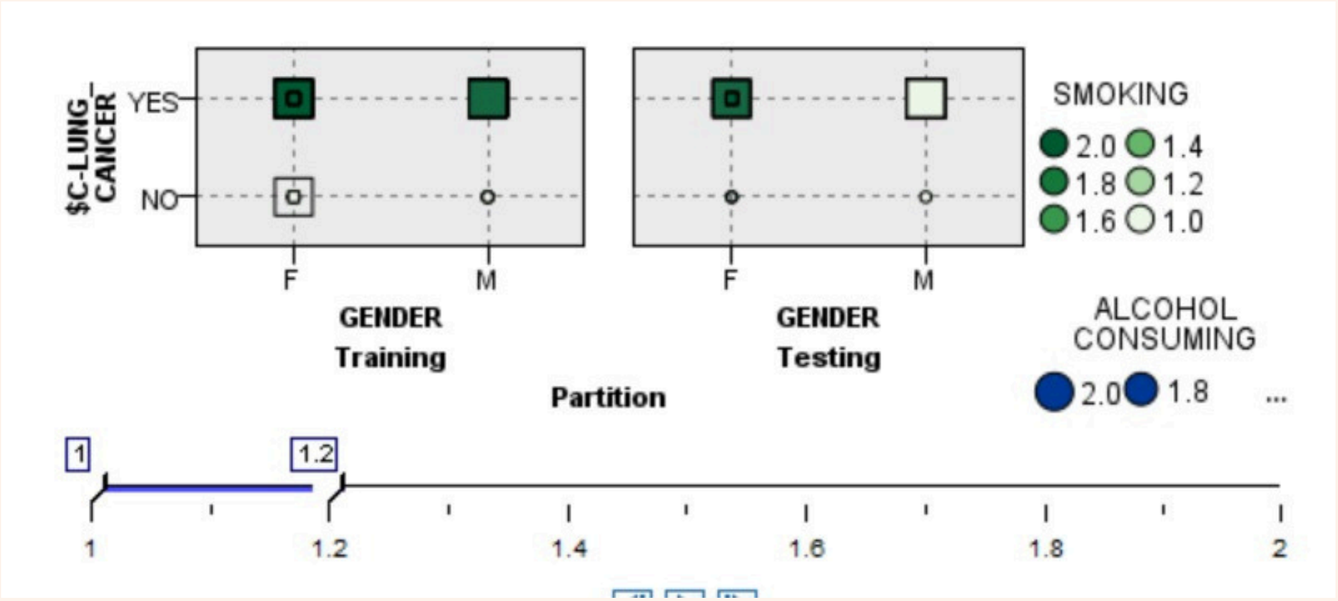
Evaluation

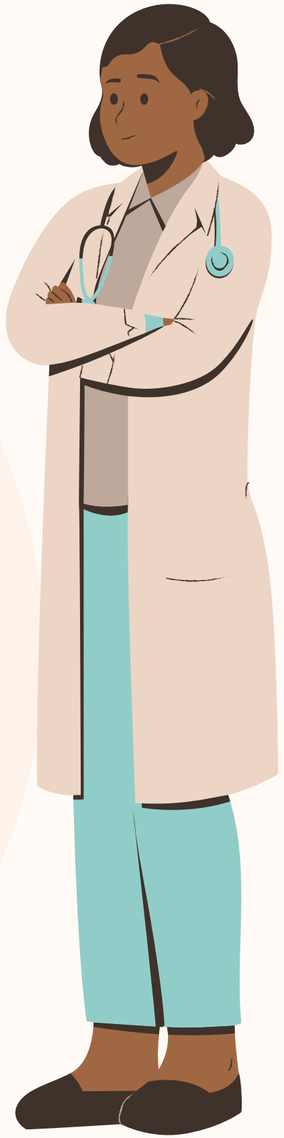
The analysis also revealed that predictors like Smoking, Yellow Fingers, Chest Pain, Anxiety, and Age were most significant in determining cancer risk. Logistic regression further provided odds ratios, quantifying how likely a patient is to develop lung cancer if specific symptoms are present. These insights help enhance medical decision-making and encourage timely diagnosis.

Our model demonstrated solid performance with an accuracy of approximately 85%, indicating that it was able to correctly classify most cases. Additionally, the ROC curve yielded an Area Under Curve (AUC) value of around 0.91, showcasing the model's strong capability to distinguish between patients with and without lung cancer.

DEPLOYMENT

Currently, the project includes a detailed report with visuals, SPSS-generated tables, and interpretations. The model logic can be shared with medical practitioners or integrated into a basic online questionnaire for use in screening patients.





Output

Results for output field LUNG_CANCER

Individual Models

Comparing $C-LUNG_CANCER with LUNG_CANCER

'Partition'	1_Training		2_Testing	
Correct	198	93.4%	88	90.72%
Wrong	14	6.6%	9	9.28%
Total	212		97	

Coincidence Matrix for $C-LUNG_CANCER (rows show actuals)

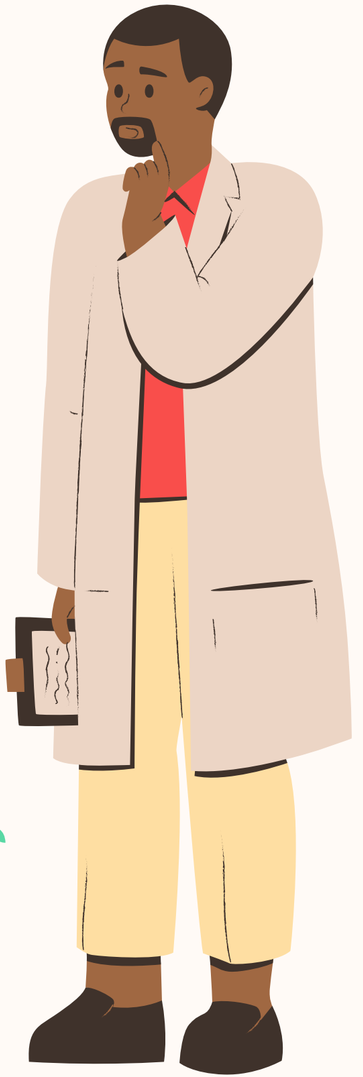
'Partition' = 1_Training	NO	YES
NO	18	9
YES	5	180
'Partition' = 2_Testing	**NO**	**YES**
NO	4	8
YES	1	84

Performance Evaluation

'Partition' = 1_Training	
NO	1.816
YES	0.087
'Partition' = 2_Testing	
NO	1.867
YES	0.041

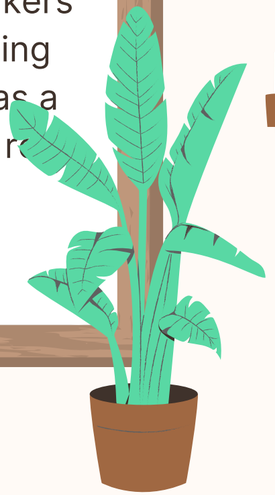
Evaluation Metrics

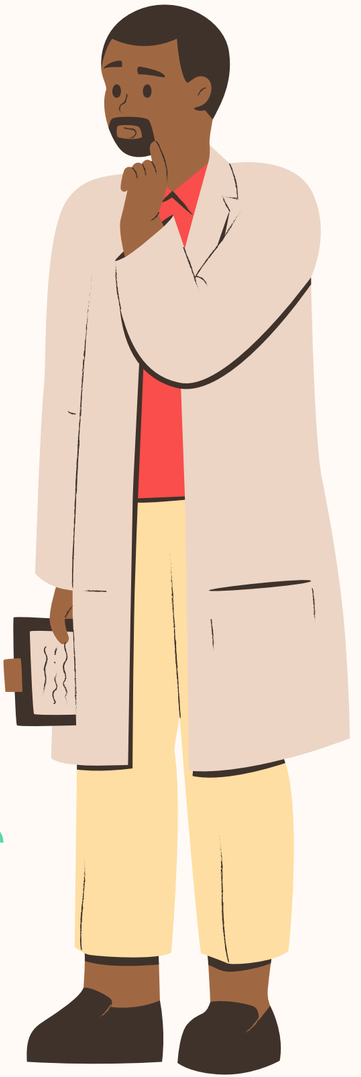
'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	G
$C-LUNG_CANCER	0.858	0.717	0.792	0.5



Applications

The lung cancer prediction model has significant real-world applications across various domains. In healthcare, it can assist doctors in early detection and clinical decision-making, especially in remote or resource-limited settings. Hospitals and diagnostic centers can integrate the model into their systems to automate preliminary screenings based on symptoms. Additionally, it can be developed into a mobile or web application, allowing individuals to self-assess their risk and seek medical advice proactively. The tool can also be valuable in public health research, helping policymakers identify at-risk populations and prioritize awareness or screening programs. Lastly, educational institutions can use this project as a case study to demonstrate the power of data science in solving r life health problems.





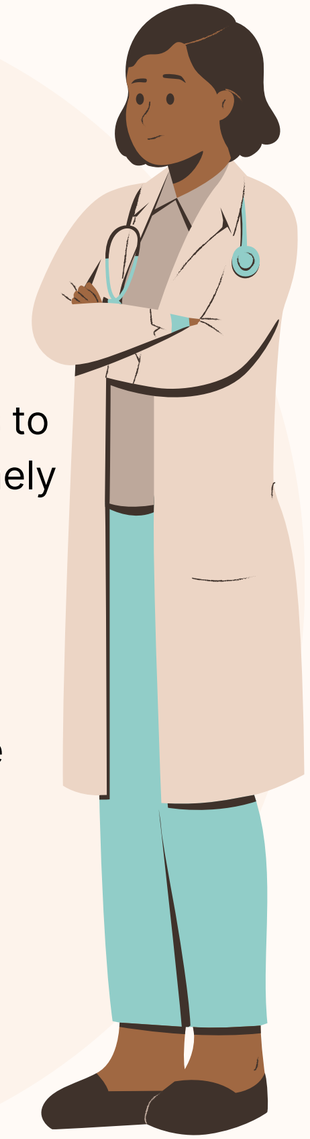
Future Scope

Integration into Healthcare Systems
The predictive model can be embedded into hospital or clinic dashboards, allowing doctors to assess patient risk in real-time using symptom-based inputs.

Mobile and Web App Development
A user-friendly app or online platform can be built where individuals input symptoms to receive instant risk assessments. This promotes self-awareness and encourages timely medical consultation.

Enhanced Model Accuracy with More Data
By collecting a larger and more diverse dataset (across regions, ages, lifestyles), the model's accuracy and generalizability can be significantly improved.

Incorporation of Medical Imaging
Future versions can include features like X-ray or CT scan image analysis using AI alongside SPSS symptom-based predictions for a more robust diagnosis.



Conclusion

Summarizing key points discussed

Our project highlights how predictive analytics using SPSS can support early detection of lung cancer. By identifying risk factors through logistic regression and validating the model's performance, we aim to contribute to public health awareness and early medical intervention. This predictive tool has potential for real-world deployment in screening systems.