Project Report

On

Lung Cancer Survival Prediction using Machine Learning

Submitted in the partial fulfillment for the award of Post Graduate Diploma in Big Data Analytics (PG-DBDA) from Know-IT ATC, CDAC ACTS, Pune

Guided by:

Mr. Milind Kapse

Submitted By:

Akash Vishwakarma (240843025005)

Swapnil Panda (240843025044)

Arbaj Riyaj Shaikh (240843025008)

Jaydeep Dhanajirao Shinde (240843025017)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Akash Vishwakarma (240843025005)

Swapnil Panda (240843025044)

Arbaj Riyaj Shaikh (240843025008)

Jaydeep Dhanajirao Shinde (240843025017)

have successfully completed their project on

Lung Cancer Survival Prediction using Machine Learning

Under the guidance of Mr. Milind Kapse

ACKNOWLEDGEMENT

This project **"Lung Cancer Survival Prediction using Machine Learning"** was a great learning experience for us and we are submitting this work to Know-IT ATC, CDAC ACTS, Pune.

We are all very glad to mention the name of **Mr. Milind Kapse** for his valuable guidance on this project. His continuous guidance and support helped us overcome various obstacles and intricacies during the course of the project work.

We are highly grateful to **Mr. Vaibhav Inamdar,** the Center Coordinator at Know-it in Pune, for his guidance and support whenever necessary while we were pursuing the Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS in Pune.

Our most heartfelt thanks goes to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) and **Mrs. Dhanshree Ma'am** (Course Coordinator, PGDBDA) who provided all the required support and his kind co-ordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

From:

Akash Vishwakarma (240843025005)

Swapnil Panda (240843025044)

Arbaj Riyaj Shaikh (240843025008)

Jaydeep Dhanajirao Shinde (240843025017)

TABLE OF CONTENTS

ABSTRACT

Lung Cancer Survival Prediction Using Machine Learning. Lung cancer is one of the leading causes of cancer-related mortality worldwide, making early and accurate survival prediction crucial for improving patient outcomes. This study explores the application of machine learning (ML) techniques to predict the survival of lung cancer patients based on clinical and demographic features. Using a dataset of 200,000 patient records, various ML algorithms including logistic regression, random forests, KNN and Extra tree classifiers and trained and evaluated to determine their predictive accuracy. Feature selection and data preprocessing techniques, such as handling missing values and feature scaling, are employed to enhance model performance. The results demonstrate that ML-based models can provide significant improvements in survival prediction compared to traditional statistical methods. This research highlights the potential of AI-driven decision support systems in oncology, enabling personalized treatment strategies and better prognostic assessments for lung cancer patients.

1. INTRODUCTION

Lung cancer is one of the most prevalent and deadliest forms of cancer worldwide, accounting for significant number of cancer-related deaths each year. Despite advancements in medical research and treatment strategies, the survival rate for lung cancer patients remains low, primarily due to late-stage diagnosis and the complexity of predicting patient outcomes. Accurate survival prediction is crucial for guiding treatment decisions, improving patient care, and optimizing healthcare resources.

Traditional survival prediction methods, such as statistical models and clinical staging systems, often struggle to capture the complex interactions between multiple risk factors, including genetic predisposition, lifestyle choices, and comorbidities. Machine learning (ML) offers a promising alternative by leveraging large datasets to identify hidden patterns and correlations that might be overlooked by conventional approaches. With the availability of extensive patien records and advanced computational techniques, ML models can enhance the accuracy of survival predictions, providing personalized prognostic insights for lung cancer patients.

This study aims to develop and evaluate ML-based models for lung cancer survival prediction using a dataset of 200,000 patient records.Various ML techniques, including supervised learning algorithms,will be explored to identify the most effective predictive model. The research will also focus on key data preprocessing steps, feature selection methods, and performance evaluation metrics to ensure robust and reliable predictions. By integrating ML into oncology, this study seeks to contribute to the development of AI-driven decision support systems that can assist healthcare professionals in making more informed and precise clinical decisions.

Datasets and features:

- Data used was collected from www.kaggle.com . These dataset provides a huge amount of information on lung cancer patients ranging from several years.
- However, overall the datasets provides a rich source of data for analyzing patterns and trends that affects survival outcomes of lung cancer patients.
- The main goal of the analysis is to build an accurate and robust regression model to predict the survival outcome of lung cancer patient. This project uses Logistic Regression, KNN ,Random Forest, Extra Tree Classifier.

2. SYSTEM REQUIREMENTS

Hardware Requirements:

- Platform – Windows 7 or above
- RAM – Recommended 8 GB of RAM
- Peripheral Devices – Keyboard, Monitor, Mouse
- WiFi connection with minimum 2 Mbps speed

Software Requirements:

- Language: Python 3
- Machine Learning
- Tableau
- OS – Windows

3. FUNCTIONAL REQUIREMENTS

1) Python 3:

- Python is a high-level programming language that is easy to learn and use.

- Python is an interpreted language, which means that code can be executed on the fly, without the need for compilation.

- Python is open source and free to use, with a large and active community of developers contributing to its development and maintenance.

- Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

2) Tableau:

- Tableau is a data visualization and business intelligence software that allows users to connect, analyse, and share data in a visual and interactive way.

- It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards, reports, and charts without the need for complex coding or programming.

- Tableau supports various data sources, including spreadsheets, databases, cloud services, and bigdata platforms, such as Hadoop and Spark.

Data Cleaning:

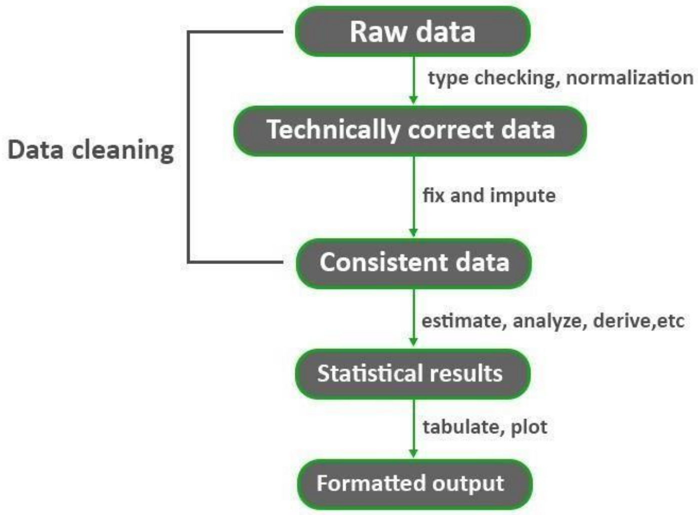

Fig: Data Cleaning Process

- Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywherein information systems. These problems are solved by data cleaning.
- Without proper data cleaning, data analysis and modelling can lead to erroneous or biased results,which can have serious consequences for businesses and organizations.
- Hence, it is a critical step in the data preparation process, as it can significantly impact the accuracyand reliability of the insights and decisions that are derived from the data. By improving the quality of data, organizations can gain a better understanding of their operations, customers, and market trends, and make more informed and effective decisions.

4. SYSTEM ARCHITECTURE

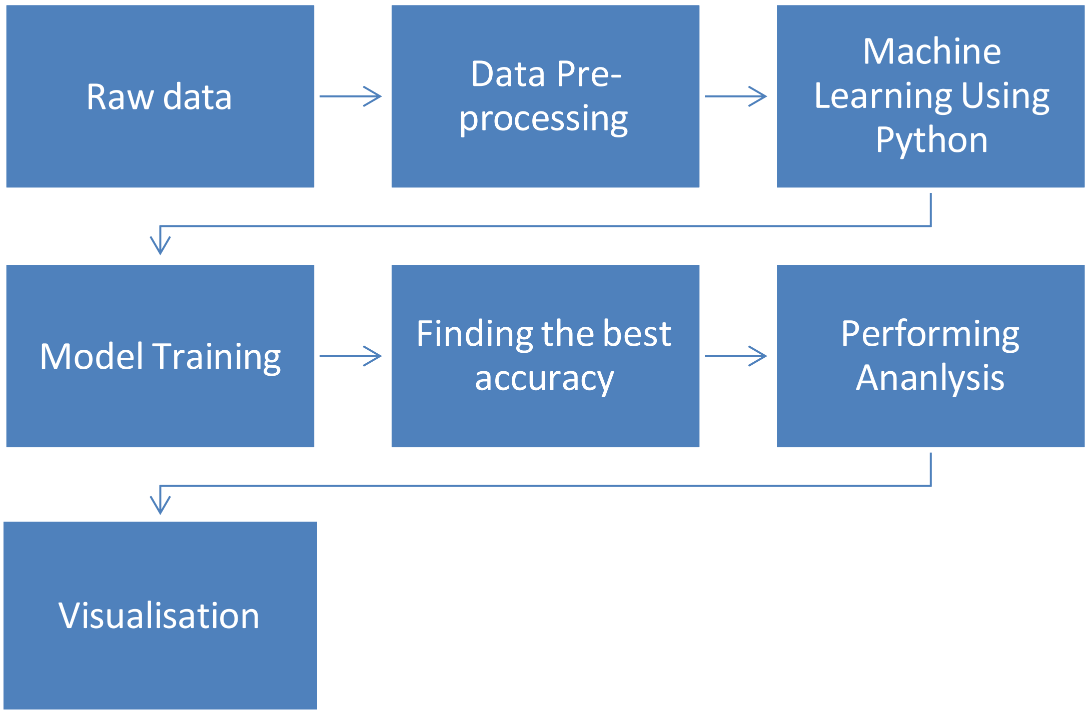

Fig: System Architecture of Lung Cancer Survival Predicition

5. METHODOLOGY

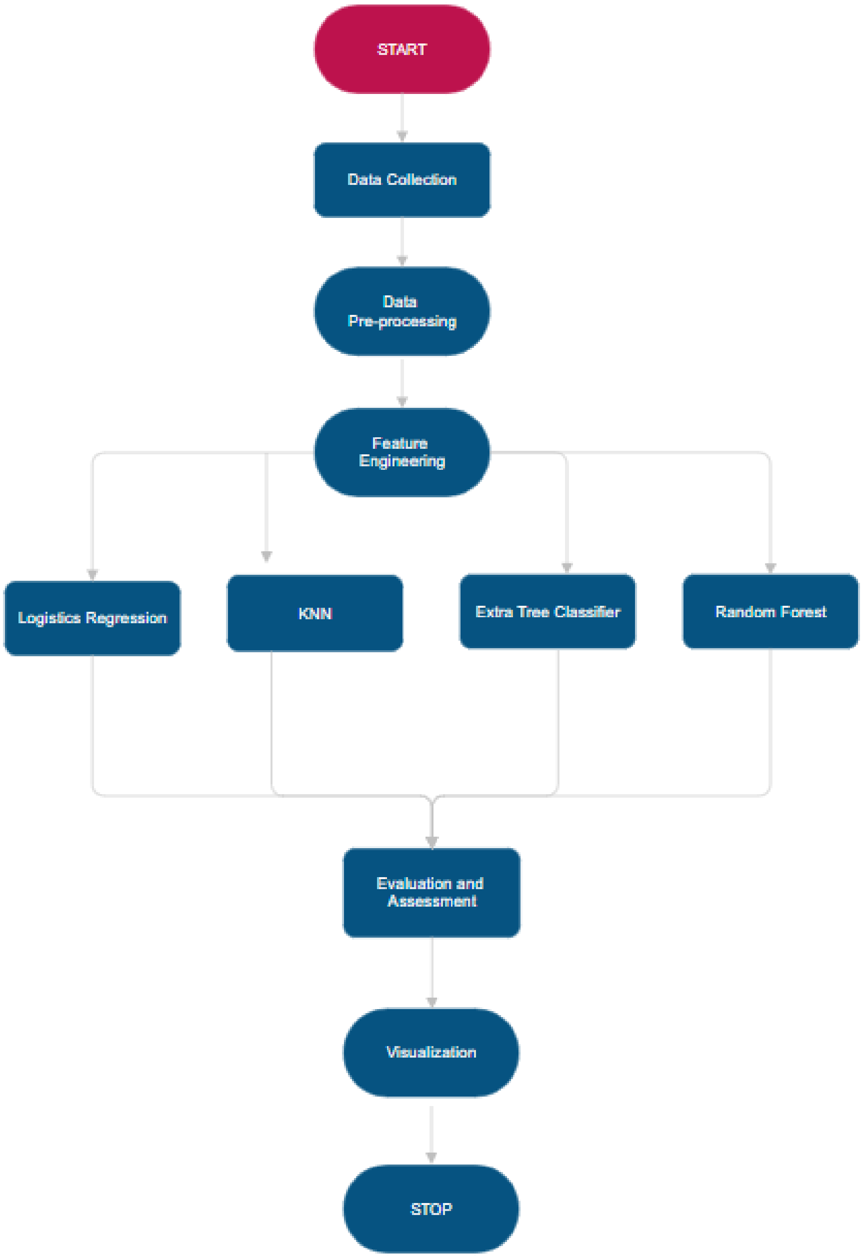

Fig: Methodology of Lung Cancer Survival Prediction

6. MACHINE LEARNING ALGORITHMS

- Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
- In our project, we applied various Regression Algorithms such as Random Forest, KNN, Logistic Regression, Polynomial Regression, and Extra Tree Classifier After the implementation, were able to analyze the accuracy of the algorithms on our data.
- Accuracy was one of the major factors that helped to decide which model has the accurate predictions.

1. Logisitc Regression

- Logistic Regression is a classification algorithm used to predict binary outcomes by estimating the probability that a given input belongs to a specific category. It works by applying the **sigmoid function** to a linear equation, transforming its output into a probability value between 0 and 1. If the probability exceeds a threshold (typically 0.5), the instance is classified into one category; otherwise, it falls into the other. The model is trained using **Maximum Likelihood Estimation (MLE)**, adjusting weights to minimize classification errors. Logistic Regression is widely used in medical diagnosis, including lung cancer survival prediction, where it helps determine the likelihood of a patient surviving based on clinical and demographic data. Its simplicity, efficiency, and interpretability make it a valuable tool in predictive analytics.

- **It starts with a linear equation:**

z=w1x1+w2x2+...+wnxn+bz = w_1x_1 + w_2x_2 + ... + w_nx_n + bz=w1x1+w2x2+...+wnxn+b
where www represents weights, xxx are input features, and bbb is the bias term.

This linear output is passed through the **sigmoid function**, which converts it into a probability between 0 and 1:
P(Y=1)=11+e−zP(Y=1) = \frac{1}{1 + e^{-z}}P(Y=1)=1+e−z1

If the probability is **greater than 0.5**, the output is classified as **1 (positive class)**; otherwise, it is **(negative class)**.

2. Random Forest:

Random forest is a machine learning algorithm that is used for classification, regression, and feature selection tasks. It is an ensemble method that combines multiple decision trees, where each tree is trained on a subset of the training data and a subset of the input features.

Pros:

- It is a highly accurate and powerful machine learning algorithm that can perform well on a wide range of classification and regression tasks.
- It can handle both categorical and continuous input variables, and it can detect and handle interactions between variables.

Cons:

- It may not perform well on small datasets or with rare or unseen classes, which may require more specialized techniques or models.
- It may not be suitable for online or real-time prediction tasks, which require faster and more lightweight models or techniques.

3. KNN Regressor:

The K-Nearest Neighbors (KNN) algorithm is a simple yet powerful classification algorithm that classifies based on a similarity measure. This supervised ML algorithm can be used for classifications and predictive regression problems1. KNN groups the data into coherent clusters or subsets and classifies the newly inputted data based on its similarity with previously trained data.

Pros:

- It is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well as regression.

Cons:

- K-NN slow algorithm: K-NN might be very easy to implement but as dataset grows efficiency or speed of algorithm declines very fast.
- Curse of Dimensionality: KNN works well with small number of input variables but as the numbers of variables grow K-NN algorithm struggles to predict the output of new data point.

4. Extra Tree Classifier

The **Extra Trees Classifier** (Extremely Randomized Trees) is an ensemble learning algorithm that improves classification accuracy by creating mul tiple decision trees and averaging their predictions.

Unlike Random Forest, which selects optimal split points based on information gain or Gini impurity, Extra Trees chooses split points **randomly**, making the model more diverse and reducing overfitting. This randomness enhances generalization and computational efficiency, as it reduces variance without significantly increasing bias. The model is trained using the entire dataset, with each tree independently learning from different random feature splits. Extra Trees is widely used in medical applications, such as lung cancer survival prediction, where it helps identify important features and improve classification performance by leveraging its robust decision-making capabilities.

Model Trained	Accuracy
Logistic Regression	47%
K-Nearest Neighbors	53%
Extra Tree Classifier	70%
Random Forest	28%

Fig. R2 Accuracy of different ML model

7. DATA VISUALIZATION AND REPRESENTATION

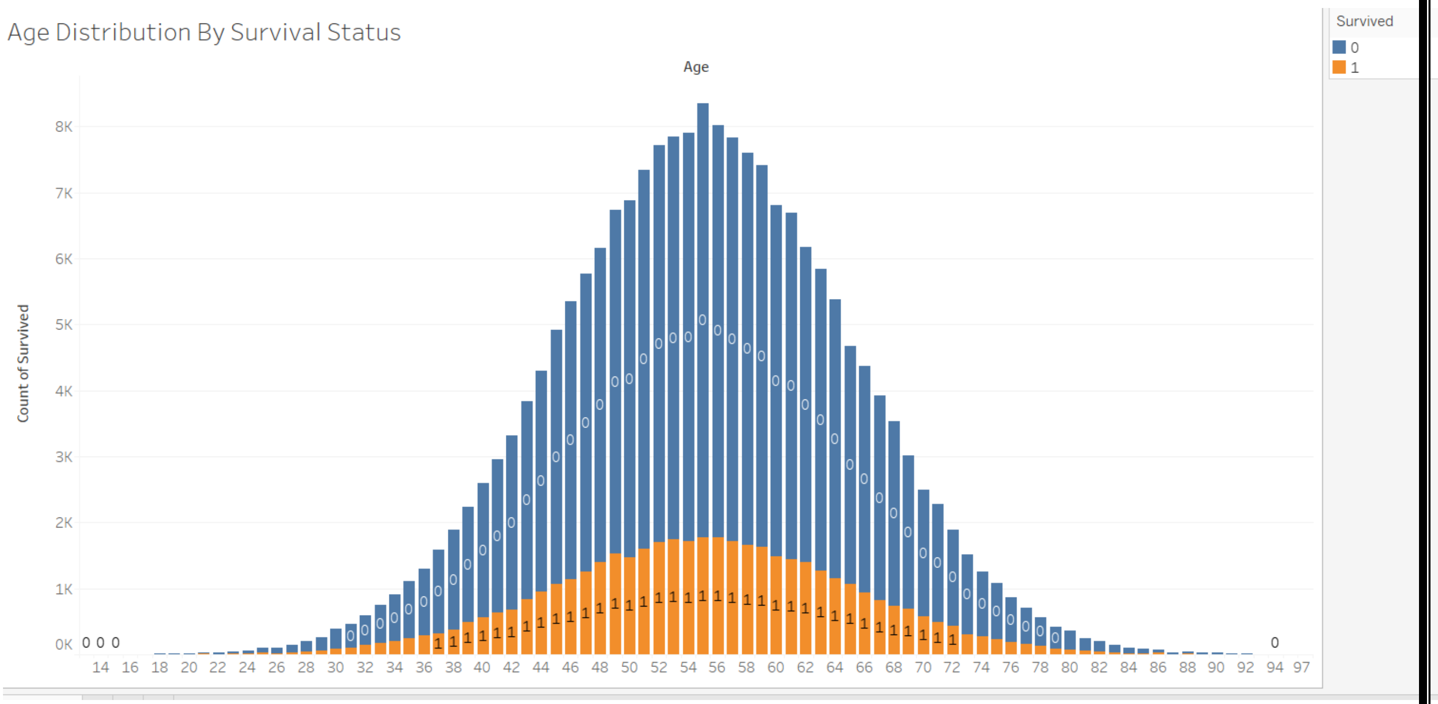

Fig: Age distribution by survival status

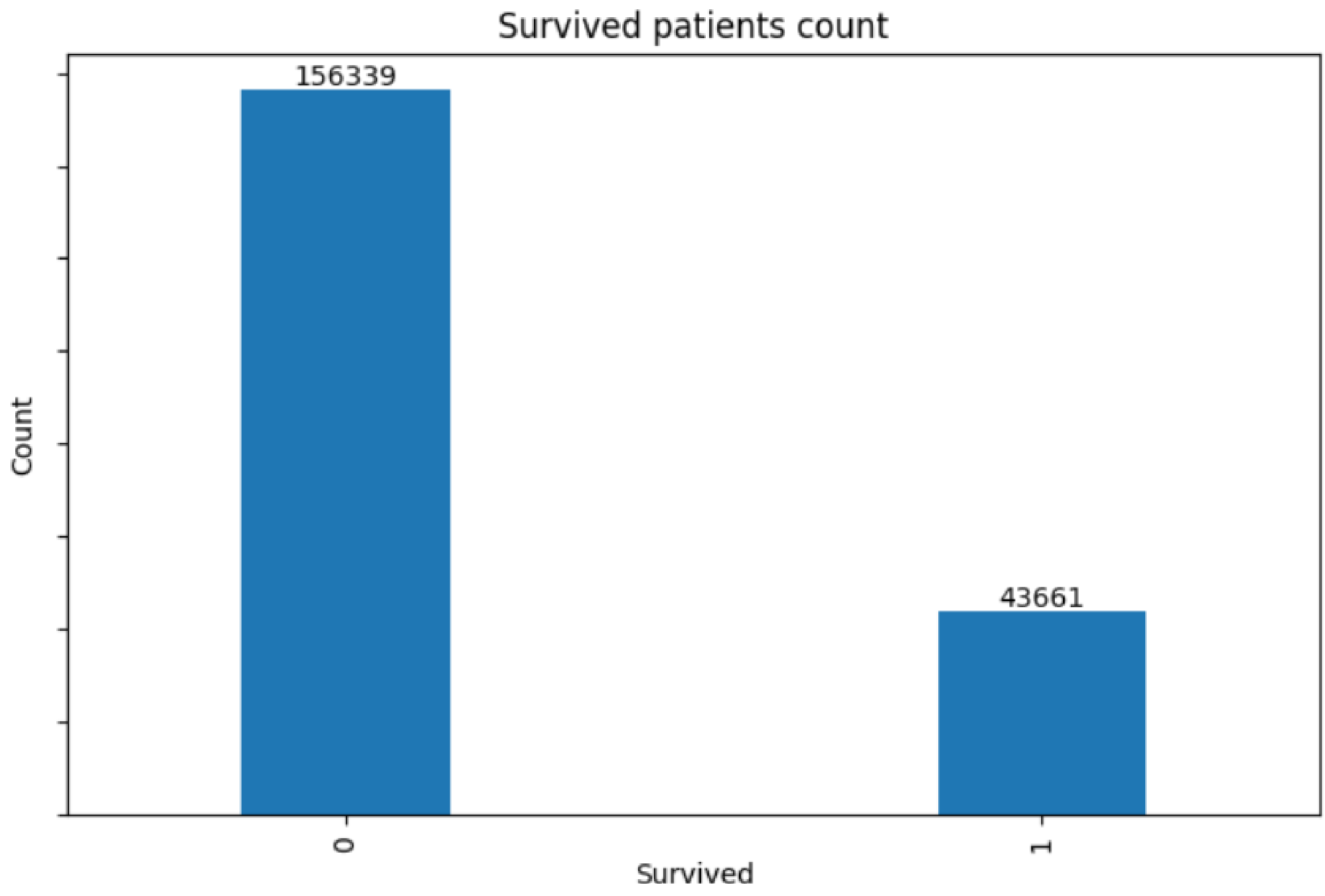

Fig. Count based on Survival status

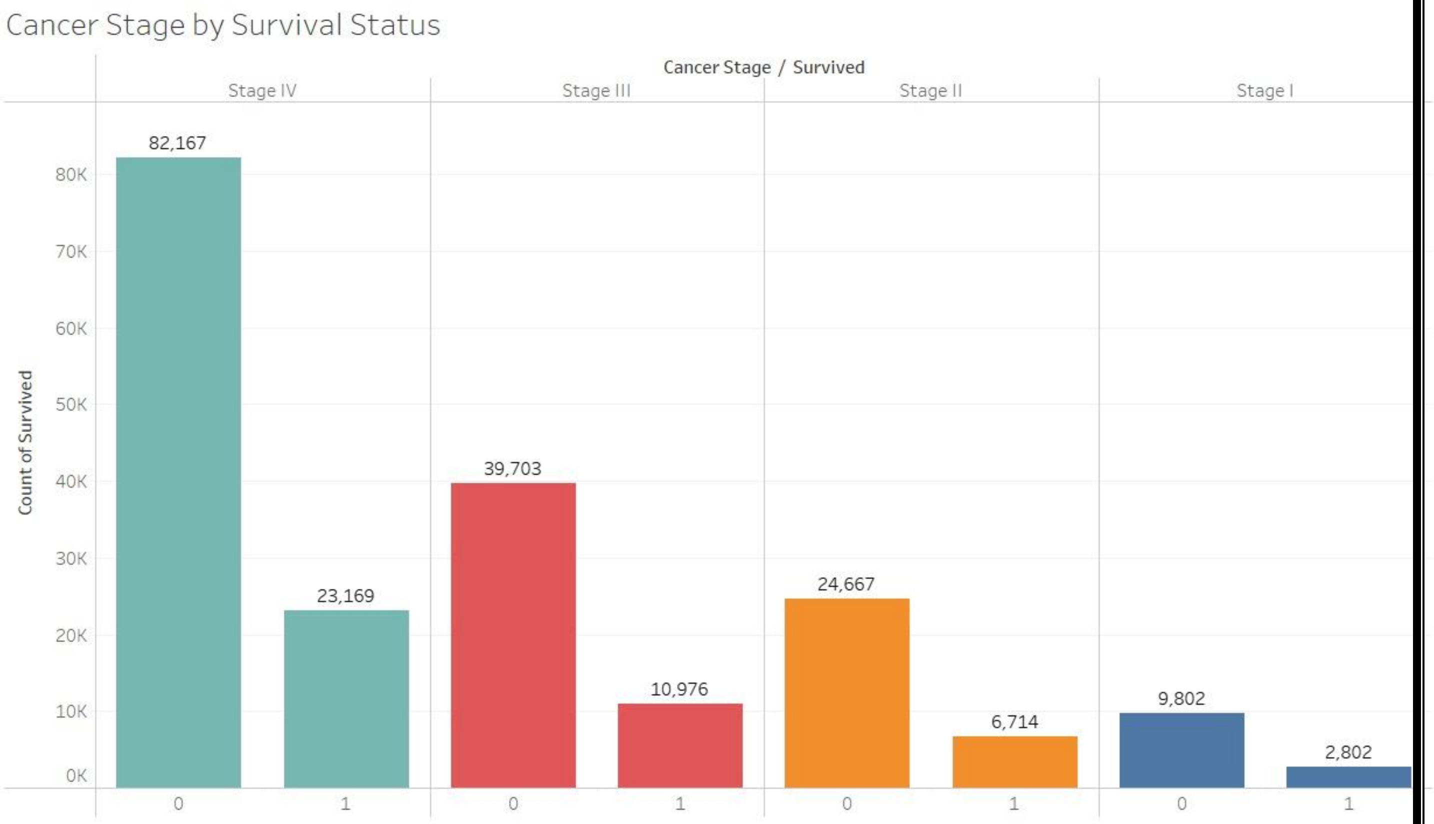

Fig. Cancer Stage by Survival Status

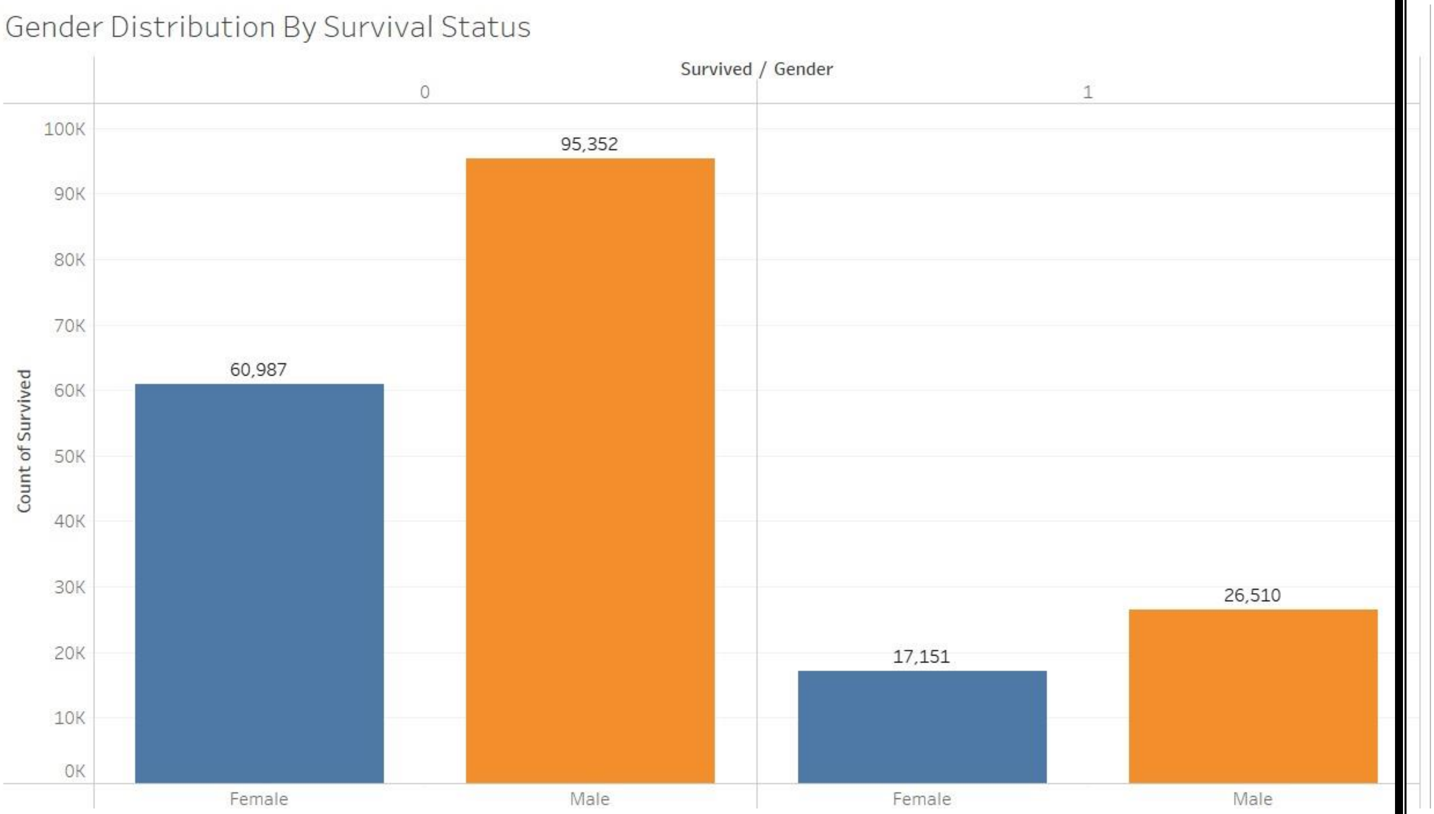

Fig. Gender Distribution by Survival Status

8. CONCLUSION AND FUTURE SCOPE

This project successfully demonstrates the application of machine learning techniques for predicting the survival of lung cancer patients. By utilizing a dataset of 200,000 patient records, some ML models,including logistic regression, random forests, KNN, and Extra Trees classifiers, were trained and evaluated. The results indicate that machine learning-based approaches provide improved predictive accuracy compared to traditional statistical methods. Data preprocessing techniques, such as handling missing values and feature scaling, played a crucial role in enhancing model performance. The findings of this study highlight the potential of AI-driven decision support systems in oncology, enabling personalized treatment plans and better prognostic assessments for lung cancer patients.

Future Scope:

Enhanced Model Performance: Further research can explore deep learning techniques, such as neural networks, to improve prediction accuracy and handle complex patterns in medical data.

Integration with Clinical Systems: Implementing the predictive model within hospital management systems can assist healthcare professionals in real-time decision-making.

Explainable AI (XAI): Future work can focus on interpretability techniques to make ML predictions more transparent and understandable for clinicians and patients.

Data Augmentation: Expanding the dataset with additional features such as genetic information, treatment history, and lifestyle factors may improve prediction reliability.

Deployment as a Web Application: Developing a user-friendly web-based or mobile application can make the predictive system accessible to doctors and patients worldwide.

Cross-validation with Multi-Center Data: Testing the model across diverse datasets from multiple healthcare institutions can validate its generalizability and robustness.

Continuous Model Updating: Implementing a continuous learning framework can help keep the model updated with new patient data, ensuring its relevance over time.

By addressing these future directions, the project can contribute significantly to the advancement of AI-driven healthcare solutions, ultimately improving patient care and survival outcomes in lung cancer treatment.

References

1. Streamlit Documentation
   https://docs.streamlit.io/

2. Dataset kaggle link
   https://www.kaggle.com/datasets/masterdatasan/lung-cancer-mortality-datasets-v2

3. Python documentation: https://docs.python.org/3/

4. "Machine Learning using Python" by Prof. U Dinesh Kumar, IIM Bangalore.

5. . Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January; I(1): 22-24.