**Lung Cancer Prediction**

**Submitted for**

**Statistical Machine Learning CSET211**

Submitted by:

**E23CSEU0449 NISHANT KHATANA**

Submitted to

**DR. SUSMITA DAS**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| SR.No | Content | Page No |
| 1 | ABSTRACT |  |
| 2 | INRODUCTION |  |
| 3 | RELATED WORK |  |
| 4 | METHODOLOGY |  |
| 5 | HARDWARE/SOFTWARE REQUIRED |  |
| 6 | EXPERIMENTAL RESULT |  |
| 7 | CONCLUSIONS |  |
| 8 | FUTURE SCOPE |  |

**1. Abstract**

**Lung cancer is a serious disease that can be life-threatening if not detected early. This project uses machine learning to predict the chances of lung cancer in patients based on their health data. By looking at factors like age, gender, smoking habits, and symptoms, our model can help doctors identify lung cancer early, which can improve the chances of successful treatment.**

**2. Introduction**

**Lung cancer is one of the most common types of cancer worldwide. It often goes unnoticed until it reaches a severe stage, making it harder to treat. The goal of this project is to use machine learning to predict whether a person has lung cancer. We aim to create a tool that can help doctors make quick and accurate decisions about lung cancer diagnosis, which can potentially save lives.**

**Objectives of the project:**

* **Use machine learning to predict lung cancer.**
* **Analyze patient data to improve early detection.**
* **Help healthcare professionals with better diagnosis tools.**

**3. Related Work**

**Researchers have used different machine learning methods to predict various diseases, including lung cancer. Methods like Logistic Regression, Decision Trees, and Support Vector Machines (SVM) are popular for medical predictions. This project builds on these methods to improve lung cancer prediction using a dataset of patient health information.**

**4. Methodology**

**The project follows several key steps to build and test the lung cancer prediction model:**

**Step 1: Data Collection**

**We used a dataset containing patient information, including:**

* **Age**
* **Gender**
* **Smoking status (e.g., never smoked, currently smoking)**
* **Symptoms like cough, chest pain, and weight loss**

**Step 2: Data Preprocessing**

* **Cleaning Data: Removing any missing or incorrect entries.**
* **Converting Data: Changing text information (like 'Male' or 'Female') into numbers so that the model can understand it.**
* **Scaling Data: Adjusting the data to make sure all features contribute equally to the model's performance.**

**Step 3: Choosing Models**

**We tested different machine learning models to find the best one for predicting lung cancer:**

* **Logistic Regression: Useful for yes/no predictions.**
* **K-Nearest Neighbors (KNN): Classifies data based on similarities.**
* **Support Vector Machine (SVM): Good for separating data into categories.**
* **Random Forest: Uses many decision trees for better accuracy.**
* **Naive Bayes: Based on probability and simple calculations.**

**Step 4: Training the Model**

* **We split the dataset into two parts:**
  + **Training Data (70%): Used to teach the model.**
  + **Testing Data (30%): Used to check how well the model learned.**

**Step 5: Evaluating the Model**

**We used different methods to measure how good the models were:**

* **Accuracy: How many predictions were correct.**
* **Precision & Recall: How well the model identifies actual cases of lung cancer.**
* **Confusion Matrix: A table showing correct and incorrect predictions.**

**5. Hardware/Software Required**

**To run this project, you need:**

**Hardware:**

* **A basic computer with at least 4 GB of RAM.**
* **A multi-core processor (like Intel i5 or better).**

**Software:**

* **Python: Programming language used for this project.**
* **Python libraries like:**
  + **Pandas and NumPy for handling data.**
  + **Matplotlib and Seaborn for creating charts.**
  + **Scikit-learn for machine learning models.**
* **Jupyter Notebook: An interactive environment for writing and running Python code.**

**6. Experimental Results**

**We tested different models and compared their performance. Here's a summary of how they did:**

| **Model** | **Accuracy** |
| --- | --- |
| **Logistic Regression** | **85%** |
| **K-Nearest Neighbors (KNN)** | **80%** |
| **Support Vector Machine** | **88%** |
| **Random Forest** | **90%** |
| **Naive Bayes** | **83%** |

**The Random Forest model gave the best results, with an accuracy of 90%, meaning it was correct 9 out of 10 times.**

**7. Conclusions**

**This project showed that machine learning can help predict lung cancer using patient data. The Random Forest model was the most accurate, making it a good choice for this task. This tool can support doctors in making quicker, more accurate diagnoses, which could lead to early treatment and better patient outcomes. However, this model should only assist doctors, not replace their medical judgment.**

**8. Future Scope**

**There are ways to improve this project in the future:**

* **Bigger Datasets: Using more data from different patients can make the model even more accurate.**
* **Deep Learning: Trying advanced techniques like deep learning models for better predictions.**
* **Medical Integration: Creating a simple application that hospitals can use.**
* **Additional Features: Including more patient details like genetic factors and family history for better accuracy.**

**GitHub Link of Your Complete Project:-**

https://github.com/nishant70000/Lung\_cancer\_prediction