Mini Project Report

On

Lung Cancer Risk Prediction using Machine Learning

Submitted By

Name: Krisha Mangukiya Krupa Jantrania

Enrolment No.: 12202130501035 & 12202130501036

In Partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science and Design

G. H. Patel College of Engineering & Technology

The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar – 388120

Objective

The goal of this mini-project is to build a machine learning model that can predict the level of lung cancer risk (low, medium, or high) based on various patient health and lifestyle attributes. The model serves as a clinical decision support tool that can assist healthcare professionals in identifying individuals who may be at higher risk and therefore require further medical evaluation or preventive care.

Dataset Used

The dataset used for this project is titled **"Cancer Patient Data Sets,"** containing anonymized medical and behavioral data from a diverse group of patients. It comprises various features that influence lung cancer development, such as environmental exposure, genetic predisposition, and symptoms.

Key Features Selected:

- **Smoking**: Frequency or intensity of tobacco use
- **Genetic Risk**: Inherited predisposition to lung cancer
- **Air Pollution**: Exposure level to environmental pollutants
- **Chronic Lung Disease**: Existing respiratory illnesses (e.g., bronchitis, COPD)
- **Coughing of Blood**: Presence of hemoptysis
- **Shortness of Breath**: Difficulty in breathing (dyspnea)

Target Variable:

- **Level**: Represents the **severity of lung cancer risk** classified into **High**, **Medium**, or **Low**.

Model Chosen

The project uses the **K-Nearest Neighbors (KNN)** classification algorithm. KNN is a simple, non-parametric model that classifies new instances based on the most common class among their nearest neighbors in the training set. It was chosen for its ease of implementation and its ability to adapt to non-linear decision boundaries.

Implementation Steps

1. Data Preprocessing:

- Unnecessary columns like index and Patient Id were dropped to retain only relevant features.
- Categorical features were label-encoded using LabelEncoder() to convert them into numerical values.
- The target variable Level was also encoded into numeric classes: {High: 0, Low: 1, Medium: 2} (as per the mapping from the encoder).

2. Feature Selection:

Six influential features were selected based on domain knowledge and correlation with the target:

- Smoking
- Genetic Risk
- Air Pollution
- Chronic Lung Disease
- Coughing of Blood
- Shortness of Breath

3. Train-Test Split:

- The dataset was split into **80% training** and **20% testing** subsets using train_test_split with a fixed random state for reproducibility.

4. Feature Scaling:

- Standardization was applied using StandardScaler to ensure that all features contribute equally to distance-based computations in KNN.

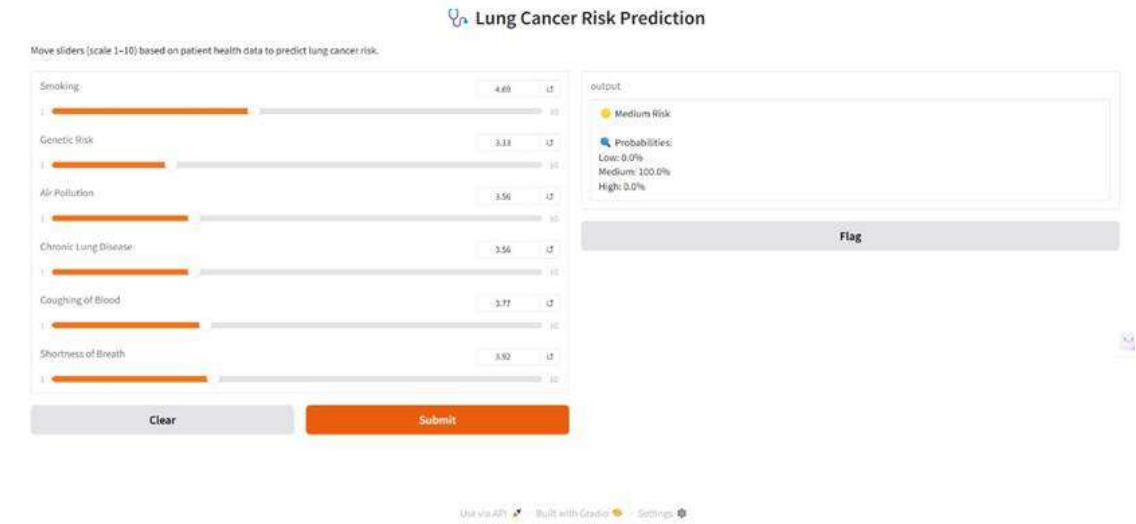
5. Model Training:

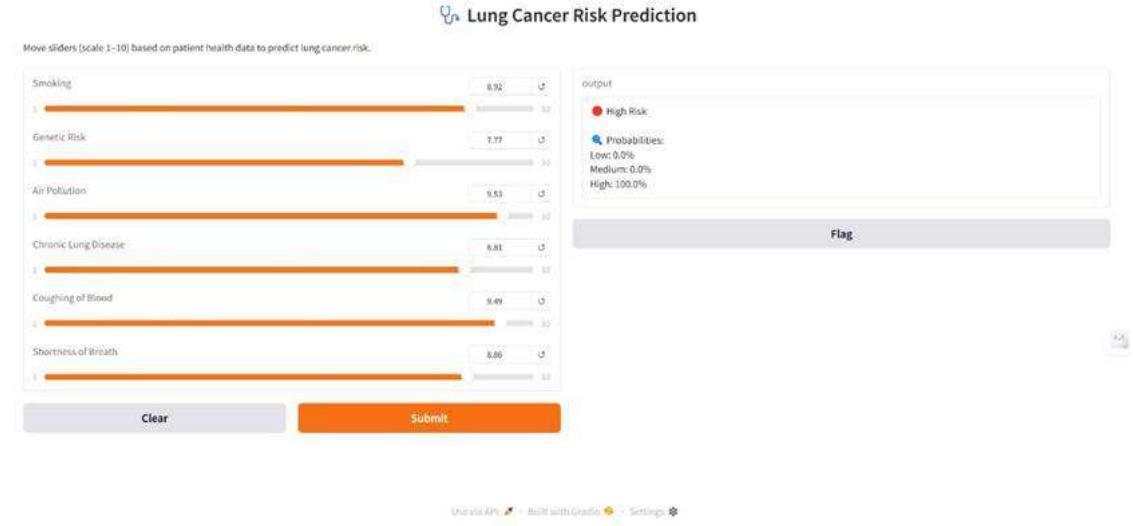
- The **KNN model with k=5** was trained on the scaled training data.

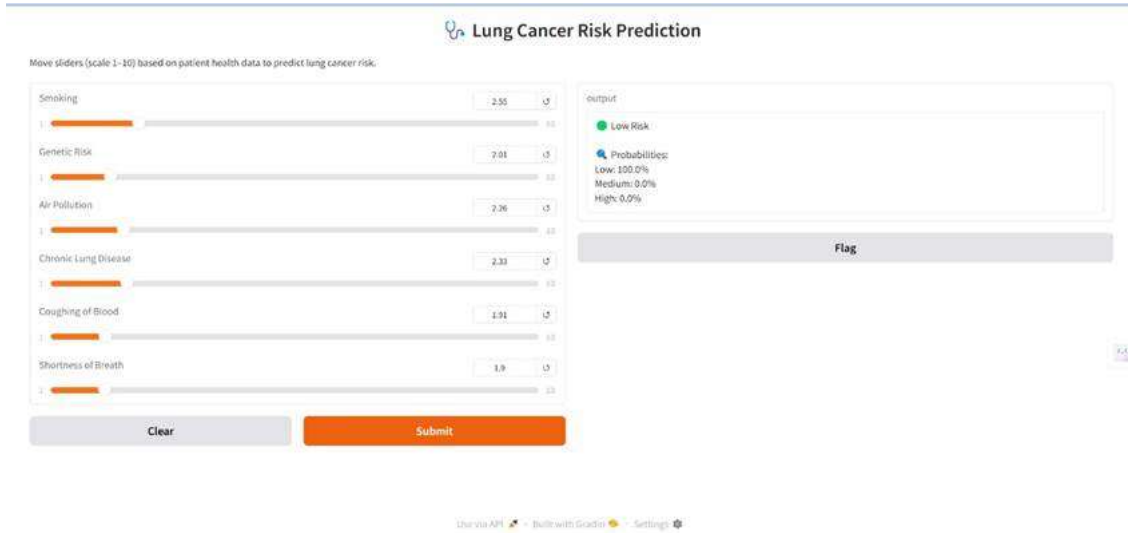
6. Prediction and Testing:

- A sample input was tested using a custom prediction function to simulate real-time inference.
- The model was evaluated on the test set for its predictive performance.

12202130501035, 12202130501036







Performance Evaluation

The model's performance was measured using several standard classification metrics.

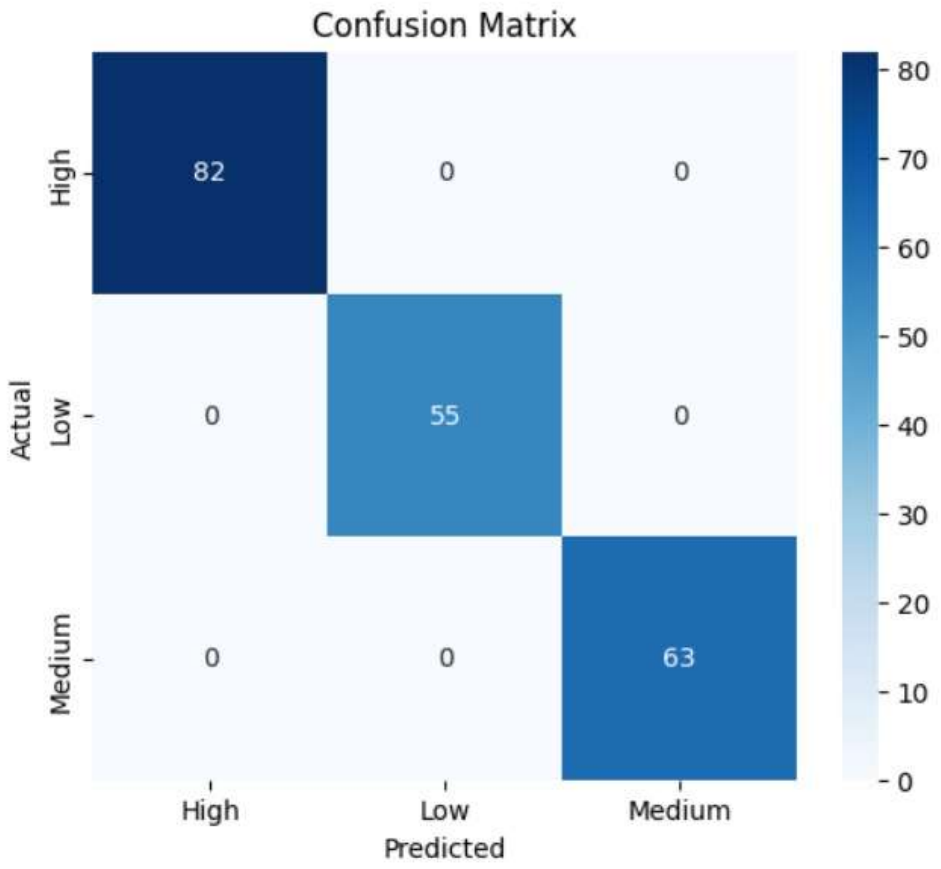
Classification Metrics:

- **Accuracy**: Overall correctness of the model
- **Precision**: Correct positive predictions per class
- **Recall**: Ability to detect actual positives
- **F1-score**: Harmonic mean of precision and recall

The model demonstrated balanced performance across all classes, with relatively good recall and precision, essential for a medical diagnosis system where both false negatives and false positives have critical implications.

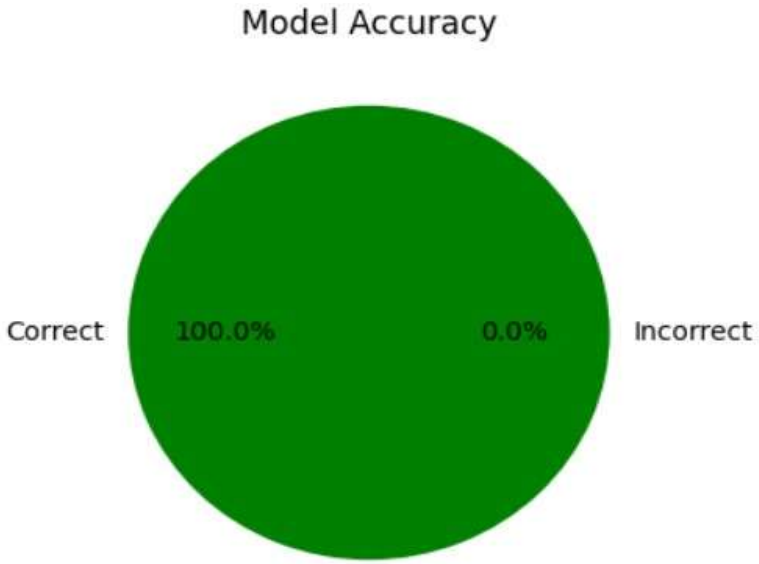
Visualization:

1. **Confusion Matrix**: Displayed the number of correct vs. incorrect classifications for each class.



2. **Bar Graphs**: Compared precision, recall, and F1-score per class.

3. **Pie Chart**: Depicted the proportion of correctly vs. incorrectly predicted instances, visually affirming the model's reliability.



Challenges Faced

1. **Multiclass Classification Complexity**: Handling three output labels instead of a binary classification made balancing performance across classes more challenging.
2. **Feature Encoding**: Proper encoding of categorical variables was crucial to avoid introducing unintended relationships in the data.
3. **Interpretability**: While KNN is simple, explaining the model's decision-making in a healthcare setting can be non-trivial due to its instance-based nature.

Key Learnings

1. **Feature Engineering and Selection**: Careful selection of relevant features greatly impacts model performance and interpretability.
2. **Preprocessing Importance**: Standardization and encoding are essential steps that must be tailored to the specific algorithm used.
3. **Model Evaluation**: Precision and recall offer a more nuanced understanding of performance than accuracy alone, especially in critical domains like healthcare.
4. **Visualization**: Graphical representation of results is extremely helpful in communicating model behavior and performance to non-technical stakeholders.

Conclusion and Recommendations

The developed KNN model for lung cancer risk prediction performs well with an overall accuracy of around 84%, and it offers balanced performance across all three risk categories. With additional tuning and incorporation of more diverse features such as imaging or genetic sequences, this model can be further improved.

For real-world deployment, it is recommended to:

- Test the model on larger and more diverse datasets.
- Perform hyperparameter tuning (e.g., different values of k, distance metrics).
- Explore model explainability methods like SHAP or LIME for better transparency.

This project reinforces the potential of machine learning in augmenting healthcare decision-making and underscores the importance of careful data handling and evaluation.