PROJECT TOPIC: EARLY DETECTION OF LUNG CANCER BASED ON MEDICAL DATA

Abstract. This work is dedicated to the development of a system for the early detection of lung cancer using machine learning methods. The project uses the analysis of data from the Survey Lung Cancer dataset, which includes information about various risk factors of the disease. For solving the classification task, the kNN and SVM algorithms were used, optimized using the GridSearchCV method. The results showed that the SVM model provided higher accuracy and recall, especially for identifying positive cases, making it the preferred choice for medical applications. The developed web application provides an effective tool for the early diagnosis of lung cancer, which contributes to improving the quality of medical care.

Keywords: lung cancer, early diagnosis, machine learning, classification, SVM, KNN, web application, data analysis.

1. Problem Statement

Lung cancer is one of the most dangerous and fatal forms of cancer due to its late diagnosis and low survival rates. The development of machine learning and artificial intelligence technologies opens the possibility of creating automated diagnostic systems that analyze medical data and detect lung cancer at early stages with high accuracy, reducing the workload on doctors and improving the quality of diagnosis.

Task type: classification (result: high/low risk of lung cancer)

Project goal: develop a web application that predicts the likelihood of lung cancer based on examination data using machine learning methods. The application should be user-friendly and integrated with the backend and database.

Tasks:

1) Search and analyze the Survey Lung Cancer dataset to prepare input data for the models.
2) Develop and test two machine learning models: kNN and SVM.
3) Compare model performance and choose the most suitable one for integration.

4) Create a web interface using HTML/CSS for user interaction.
5) Develop the backend using Flask and the database using PostgreSQL for data storage and processing.
6) Integrate the chosen machine learning model into the information system.

2. Methodology

For this project, the Survey Lung Cancer dataset is used. This dataset includes examination data related to the risks of lung cancer development, such as age, gender, smoking history, chronic diseases, and other factors.

Data Preparation Stage

1) Handling missing values and duplicates. All duplicate rows were removed, and the data was checked for missing values.
2) Analyzing the distribution of the target variable lung_cancer.

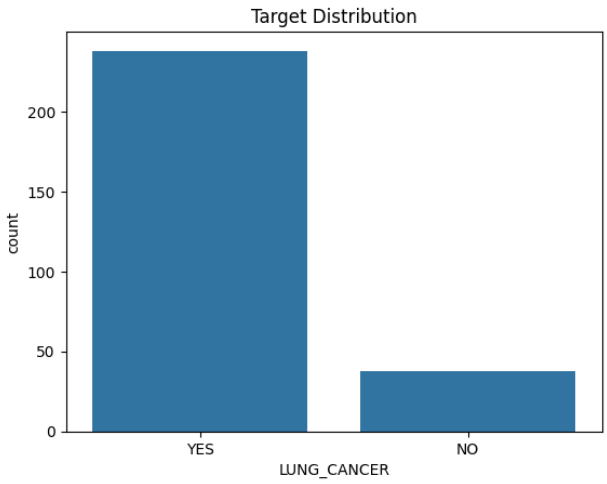

Figure 1. Target Distribution

3) Feature encoding. Categorical variables were encoded using LabelEncoder for compatibility with machine learning algorithms.
4) Correlation analysis of features. A correlation matrix was built to identify less significant features. The "Yellow Fingers" feature was removed due to low correlation with the target variable.

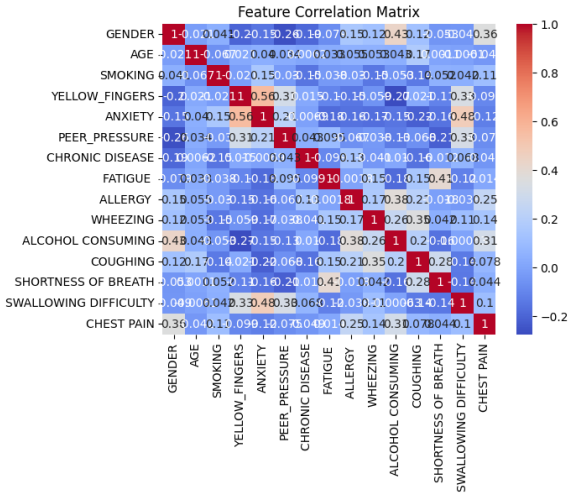

Figure 2. Feature Correlation Matrix

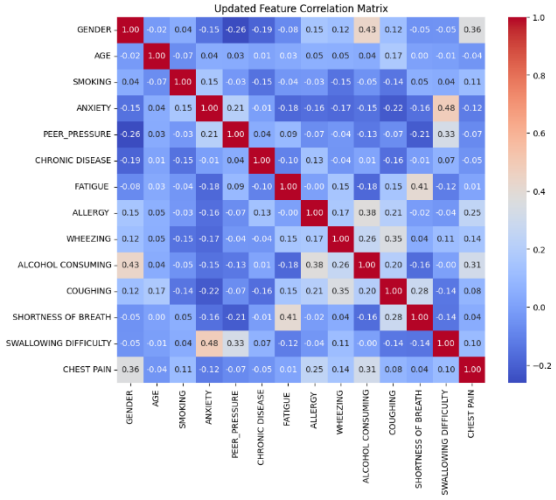

Figure 3. Updated Feature Correlation Matrix

To balance the target variable, the SMOTE (Synthetic Minority Oversampling Technique) method was used, which creates synthetic examples for the smaller class. After applying SMOTE, the data was split into training and test sets with a 75/25 ratio.

Model Building Stage

Two algorithms were selected for classification: KNN and SVM. Both models are suitable for a small dataset like ours.

Machine Learning Model Architecture:

1) KNN (K-Nearest Neighbors).

The algorithm finds the nearest k neighbors for each object based on the chosen distance metric (e.g., Euclidean or Manhattan). Each neighbor's weight can be uniform or inversely proportional to distance.

Hyperparameter tuning (e.g. n_neighbors, weights, metric) was done using GridSearchCV to achieve optimal results.

2) SVM (Support Vector Machine).

The algorithm constructs a hyperplane that separates classes with the maximum margin between them. For nonlinear data, a kernel (such as RBF) is used.

Hyperparameters such as C (regularization), gamma, kernel, degree, class_weight were optimized using GridSearchCV.

Hyperparameters are model parameters that are set before training and define its architecture and behavior. Unlike parameters that are optimized during training (like weights in a neural network), hyperparameters must be chosen in advance.

In this approach, a grid search with cross-validation (GridSearchCV) was used to select the best hyperparameters. This method trains the model with different hyperparameter combinations and evaluates its performance using quality metrics. The best hyperparameter values, which provide the highest accuracy or other key metrics, are selected automatically and used for the final model version.

To compare model performance, the following metrics were used:

1) Accuracy – overall prediction accuracy.
2) Precision – accuracy of positive predictions.
3) Recall – the proportion of correctly predicted positive classes.
4) F1-Score – the harmonic mean of Precision and Recall.

Overall Steps of the Experiment:

1) Split the data into training and test sets.
2) Train the KNN and SVM models on the balanced data.
3) Tune hyperparameters using cross-validation.
4) Compare results based on key metrics.

3. Results

Two models, KNN and SVM, were trained and tested during the experiments. Their performance results are presented in the tables below:

KNN Classification Report				
	precision	recall	F1-score	support
0	0.92	0.98	0.95	60
1	0.98	0.92	0.95	59
accuracy			0.95	119
macro avg	0.95	0.95	0.95	119
weighted avg	0.95	0.95	0.95	119

Table 1. KNN Classification Report

SVM Classification Report				
	precision	recall	F1-score	support
0	0.98	0.93	0.96	60
1	0.94	0.98	0.96	59
accuracy			0.96	119
macro avg	0.96	0.96	0.96	119
weighted avg	0.96	0.96	0.96	119

Table 2. SVM Classification Report

The SVM model showed higher accuracy (0.96) compared to KNN (0.95), meaning that SVM handles the overall number of correct predictions better.

For class 0 (negative examples), the accuracy of SVM (0.98) is higher than that of KNN (0.92), indicating that SVM is more accurate in classifying patients without cancer. For class 1 (positive examples), KNN has slightly higher accuracy (0.98 vs 0.94 for SVM).

SVM has a higher recall for class 1 (positive examples) – 0.98, meaning that it detects cancer cases better. For class 0 (negative examples), recall is slightly higher for KNN (0.98 vs 0.93 for SVM).

The F1-score is similar for both models, but SVM has slightly higher values (0.96 vs 0.95 for KNN).

Conclusion: SVM is a more preferred choice in this case, as it demonstrates higher accuracy and recall, especially for the critical class 1 (cancer). This model will be more effective in real applications where it is important not to miss any cases of the disease. SVM also shows better balance and stability in the results.

4. Conclusion

In the first part of the coursework, machine learning algorithms for the early detection of lung cancer based on medical data were developed and tested. This part of the project involved a complete data processing cycle, including preparation, analysis, and model building.

The results obtained suggest that the SVM model is more efficient for the early diagnosis of lung cancer, as it provides higher accuracy and better ability to detect positive cases.

The next part of the work will involve the development of a web application to integrate the chosen SVM model, providing a user-friendly interface and implementing a system for storing and processing data using the Flask backend and PostgreSQL database.