CM 3015 Machine Learning and Neural Networks

Midterm Coursework Assignment

Report

1. Abstract

This project undertakes a comprehensive comparison of three distinct machine learning models—K-Nearest Neighbors, Logistic Regression, and Gaussian Naive Bayes—within the context of predicting lung cancer based on a set of symptoms. The primary objective is to evaluate and understand the strengths and weaknesses of each model, contributing valuable insights to the broader landscape of machine learning applications. Through systematic training, evaluation, and analysis, this report presents a nuanced exploration of model performance, ultimately identifying Logistic Regression as the most promising model for the given dataset.

2. Introduction

In the realm of machine learning, the selection of an appropriate model is pivotal for achieving accurate and reliable predictions. This project focuses on three well-established models: K-Nearest Neighbors, Logistic Regression, and Gaussian Naive Bayes. The aim is to assess their performance on a dataset specifically designed for predicting lung cancer based on a set of symptoms. This investigation is motivated by the need to understand how each model copes with the complexities inherent in the dataset, thereby informing their practical applicability.

The dataset employed in this project consists of 272 entries, each characterized by 11 features. These features include symptoms such as yellow fingers, peer pressure, chronic diseases, fatigue, allergy, wheezing, alcohol consumption, coughing, swallowing difficulty, and chest pain. The target variable is 'LUNG_CANCER,' indicating the presence or absence of lung cancer. All features are binary, and the dataset is initially inspected to ensure completeness and appropriateness for the machine learning task.

Lung cancer prediction is a critical task with significant implications for healthcare. Accurate and timely identification of potential cases can facilitate early intervention and improve patient outcomes. The selection of an optimal machine learning model is crucial for creating a reliable system capable of assisting individuals in assessing their risk of developing lung cancer based on observable symptoms.

The project adopts a systematic approach, encompassing data preprocessing, model training, evaluation, and further analysis. Three models are considered, each undergoing rigorous evaluation using common metrics such as accuracy, precision, recall, and F1-score. Cross-validation is employed to assess model stability, and a detailed classification report is generated for each model.

3. Background

3.1 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric and lazy learning algorithm utilized for both classification and regression tasks. The core principle involves making predictions based on the majority class or mean value of its k-nearest neighbors in the feature space.

My code initializes a KNN model, fits it to the training data, predicts on the test set, and evaluates its performance using accuracy. Cross-validation is employed to assess the model's stability and generalization across different folds.

3.2 Logistic Regression

Logistic Regression, despite its name, is a linear model primarily used for binary classification tasks. The algorithm predicts the probability that an instance belongs to a particular class using the logistic function.

The Logistic Regression model is instantiated, trained on the training set, and evaluated on the test set in the provided code. Cross-validation is performed to gauge the model's consistency across different subsets.

3.3 Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic model grounded in Bayes' theorem. It assumes that features are conditionally independent given the class label, simplifying the computation of probabilities.

The code provides a custom implementation of Gaussian Naive Bayes, covering training, prediction, and evaluation. Cross-validation scores are computed to assess the model's consistency across different folds.

4. Methodology

The exploration of the dataset and the application of machine learning algorithms involved a systematic approach to ensure robust model evaluation and comparison.

The initial step included a thorough examination of the dataset, named 'modelData,' which comprised 272 entries and 11 columns. The last column, 'LUNG_CANCER,' served as the target variable. All columns had non-null integer values, and the dataset was free of missing data.

4.2 Feature Extraction and Target Extraction

The features matrix (X) was created by excluding the target variable ('LUNG_CANCER') from the 'modelData' DataFrame. This matrix represented the input features for model training.

The target array (y) was extracted by selecting only the 'LUNG_CANCER' column from the 'modelData' DataFrame. This array represented the ground truth labels for the corresponding features.

4.3 Data Splitting

The dataset was split into training and test sets using the train_test_split function from scikit-learn. The split was performed with a test size of 30% and a random seed (random_state=42) for reproducibility. This partitioning ensured that a portion of the data was reserved for evaluating the models' performance independently.

4.4 Machine Learning Models

Three machine learning models were selected for evaluation: K-Nearest Neighbors (KNN), Logistic Regression, and Gaussian Naive Bayes.

4.4.1 Model Training

The models were instantiated and trained on the training set (Xtrain, ytrain).

4.4.2 Prediction

The trained models were used to predict the target values on the test set (Xtest), assessing their generalization performance.

4.4.3 Model Evaluation

The accuracy of each model was calculated using the accuracy_score function, providing a quantitative measure of their predictive performance. Additionally, classification reports were generated to obtain precision, recall, and F1-score for each class.

4.5 Cross-Validation

Cross-validation was employed as a crucial technique for assessing the models' stability and generalization across different subsets of the dataset. Cross-validation scores were computed using the cross_val_score function with 5 folds for both K-Nearest Neighbors and Gaussian Naive Bayes models. For Logistic Regression, cross-validation scores were obtained separately.

4.6 Algorithm Modifications and Exploration

4.6.1 Gaussian Naive Bayes Custom Implementation

A custom implementation of Gaussian Naive Bayes was developed to enhance understanding and allow for flexibility in adjusting the algorithm. This included a class (GaussianNaiveBayes) containing methods for training, prediction, and evaluation.

4.6.2 Logistic Regression Hyperparameter Exploration

The Logistic Regression model's behaviour was further explored by investigating the impact of the regularization parameter (C) on both training and validation accuracy. A validation curve was generated using the validation curve function, plotting the training and validation scores across a range of C values.

4.7 Model Comparison and Selection

The performance metrics obtained from the evaluation and cross-validation were systematically compared for each model. Logistic Regression emerged as the most promising model based on its consistently high accuracy, balanced performance across classes, and favourable cross-validation scores.

4.8 Further Analysis on Logistic Regression Model

A deeper analysis of the Logistic Regression model included:

4.8.1 Validation Curve

A validation curve was plotted to visualize the impact of different regularization parameter values on training and validation accuracy. This analysis aimed to identify an optimal range for the regularization parameter.

4.8.2 Model Evaluation Metrics

Various evaluation metrics were calculated for the Logistic Regression model, including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provided a comprehensive understanding of the model's predictive capabilities and highlighted its strengths and areas for improvement.

In summary, the methodology involved meticulous data exploration, proper data splitting, training and evaluating multiple machines learning models, leveraging cross-validation for robust assessment, and conducting additional analyses for a deeper understanding of the Logistic Regression model's behaviour. This systematic approach ensured the

reliability and validity of the findings, ultimately guiding the selection of the most suitable model for predicting lung cancer based on the provided symptoms.

5. Results

ML Algorithms	Model Accuracy	Cross-Validation
K-Nearest Neighbors Model	86.59%	80.0% to 92.59%.
Gaussian Naive Bayes Model	89.02%	85.45% to 87.27%.
Logistic Regression	90.24%	87.27% to 92.73%.

Logistic Regression stands out with balanced performance for both classes, aligning with the project's objective.

6. Evaluation

6.1 Achievements

6.1.1 Model Performance

The project successfully achieved its primary aim of systematically comparing the performance of three machine learning algorithms—K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Logistic Regression—for predicting lung cancer based on provided symptoms. The Logistic Regression model emerged as the most promising, achieving the highest accuracy (90.24%) and demonstrating balanced performance across classes.

6.1.2 Cross-Validation and Stability

Cross-validation was employed to assess the stability and consistency of each model. Logistic Regression consistently performed well across different folds, reinforcing its reliability and suitability for diverse subsets of the dataset.

6.1.3 Classification Reports

Detailed classification reports provided insights into the strengths and weaknesses of each model. The Logistic Regression model exhibited balanced precision, recall, and F1-score for both classes, aligning with the project's objective of creating a model applicable to diverse scenarios.

6.1.4 Further Analysis

The project delved into additional analysis, including a validation curve for Logistic Regression. This analysis contributed to a deeper understanding of the impact of regularization on the model's performance.

6.2 Challenges and Limitations

6.2.1 Class Imbalance

The dataset exhibited class imbalance, with a limited number of positive cases (lung cancer patients). This imbalance impacted the performance metrics, especially for class 0 in the classification reports. Strategies such as oversampling or adjusting class weights could be explored to address this issue.

6.2.2 Limited Feature Exploration

The project used a specific set of symptoms as features for predicting lung cancer. Further exploration of additional symptoms, demographic factors, or interactions between features could enhance the model's predictive capabilities. However, this would require access to a more comprehensive dataset.

6.2.3 Hyperparameter Tuning

The project did not extensively explore hyperparameter tuning for each model. Fine-tuning hyperparameters, especially for complex models like KNN or Logistic Regression, could potentially lead to further improvements in performance.

6.3 Critical Reflection

6.3.1 Research Aim

The project's aim was to compare machine learning models for lung cancer prediction. While the aim was achieved, the complexity of lung cancer prediction and the limitations of the dataset may prevent achieving a highly accurate predictive model within the scope of this project.

6.3.2 Real-world Applicability

The project primarily focused on model performance within the given dataset. Real-world applicability might require additional considerations, such as ethical implications, interpretability of models, and the impact of false positives or false negatives on individuals.

6.4 Future Directions

6.4.1 Data Expansion

Collecting a more extensive and diverse dataset with a balanced distribution of classes would improve the robustness and generalization of the models.

6.4.2 Feature Engineering

Exploring additional symptoms, lifestyle factors, or genetic information could contribute to a more comprehensive understanding of lung cancer risk.

7. Conclusions

In conclusion, this project aimed to systematically compare the performance of three machine learning algorithms—K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Logistic Regression—for predicting lung cancer based on provided symptoms. The findings reveal that Logistic Regression outperformed the other models, achieving the highest accuracy of 90.24%.

The classification reports highlighted the balanced performance of Logistic Regression across both classes, emphasizing its suitability for diverse scenarios.