A Novel Machine Learning Model Using Ensemble Learning and High-Performance Filter for Lung Cancer Prediction

by

Group No. 02
Kazi Sati (Exam roll: 192291, Class roll: 1974)
Pinky Akter (Exam roll: 192294, Class roll: 1977)
Md. Zuleyenine Ibne Noman (Exam roll: 192319, Class roll: 2002)

A Project Report submitted to the
Institute of Information Technology
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Information Technology

Md. Mahmudur Rahman
Lecturer
Institute of Information Technology
Jahangirnagar University

Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342
October, 2023

DECLARATION

We hereby declare that this project is based on the results found by ourselves. Materials of work found by other researcher are mentioned by reference. This project, neither in whole nor in part, has been previously submitted for any degree.

Roll:192291

Roll:192294

Roll:192319

CERTIFICATE

The project titled "A Novel Machine Learning Model Using Ensemble Learning and High-Performance Filter for Lung Cancer Prediction" submitted by Kazi Sati (ID: 1974, Session: 2018-19), Pinky Akter (ID: 1977, Session: 2018-19), Md. Zuleyenine Ibne Noman (ID: 2002, Session: 2018-19) has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Information Technology on the 1st of Novemver 2023.

Md. Mahmudur Rahman
Lecturer
Institute of Information Technology
Jahangirnagar University

ABSTRACT

Lung cancer is the deadliest among all diseases. This project aims to develop a robust model to predict lung cancer based on given features in a dataset using machine learning techniques. The model is prepared to receive patient/people information as a single csv (comma separated values) file. The dataset is prepared by handling null values, duplicates, using feature engineering after data analysis and performing scaling. Later the preprocessed data is fed to six machine learning algorithms, namely, Logistic Regression (LR), Decision Tree classifier (DT), Random Forest classifier (RF), K-Nearest Neighbors classifier (KNN), Gaussian Naive Bayes (GNB), Support Vector Classifier (SVC). Based on their performance, the top three algorithms are selected for being used as estimators in the Ensemble learning process. Cross-validation is used to prevent biased results. Hyperparameter tuning is used, namely grid search to find the optimal number of neighbors in KNN algorithm. We used four datasets of different sizes and different nature. The model is run on each of these dataset for having broader insight of its performance.

Keywords: Ensemble Learning, Cross-validation, Logistic Regression, Decision Tree classifier, Random Forest classifier, K-Nearest Neighbors classifier, Gaussian Naive Bayes, Support Vector Classifier.

LIST OF ABBREVIATIONS

LR	Logistic Regression
DT	Decision Tree
Rf	Random Forest
KNN	K-Nearest Neighbors
GNB	Gaussian Naive Bayes
SVC	Support Vector Classifier
ML	Machine Learning

LIST OF FIGURES

Figure

LIST OF TABLES

Table

TABLE OF CONTENTS

CHAPTER I

Introduction

1.1 Overview

Being the deadliest disease, lung cancer causes one-third deaths of all cancers. But if lung cancer can be diagnosed at an early stage, the chance of survival will be greatly increased. [1] Therefore, predicting lung cancer accurately is important. Machine learning can detect patterns and learn to predict lung cancer with greater accuracy. Our model predicts lung cancer using mostly categorical matrics. Dataset plays a significant role for accurate prediction. Selecting a non-biased and larger dataset is important. In kaggle, we have searched for datasets of CSV format that are on Lung Cancer. We come across four different datasets. Our model receives a csv file from each dataset. We first predict lung cancer using Logistic regression, Decision Tree algorithm, K-Nearest Neighbor, Random Forest Tree, Support Vector Machine, Naive Bayes algorithm separately. The model collects the top performance algorithms in an optimal way such that maximizes accuracy while keeping computational complexity as low as possible. Then using ensemble learning the model can predict lung cancer more accurately.

1.2 Problem Statement

One-third deaths of all cancers are caused because of lung cancer.[1] If lung cancer can be diagnosed at an early stage, the chance of survival will be greatly increased. [2] Therefore, predicting lung cancer accurately is important.Machine learning can detect patterns and learn to predict lung cancer with greater accuracy. This model predicts lung cancer using mostly categorical matrics. Machine learning model can perform at level of an expert. But, being expert in real-life requires time and experience. Also,

human is prone to error. Therefore, implementing a model with machine learning is essential for having sufficient support for healthcare.

1.3 Motivation

Machine learning can come to human aid by providing highly accurate prediction about diseases like lung cancer with minimal cost, being 24 hour available. High accuracy of a model depends on quality and quantity of data, feature selection, data processing, model selection, hyper tuning, cross-validation, model not over-fitting under fitting, model ensembling. To the best of our knowledge, we have seen different algorithms having different strengths of predicting correctly some features while having weakness in finding others. None of these works combine the strengths which can make the model more generalized and robust. Hence, a hybrid machine learning model can predict better about cancer with higher accuracy for data with different patterns.

1.4 Objective

The main objective of this project is to build a robust model by combining best performing machine learning algorithms, while reducing complexity as much as possible, also by preventing biasness. Specific goals of this thesis that should be mentioned:

- Combine strength of different machine learning models for higher performance.
- Utilizing cross-validation is to assess a machine learning model's performance.
- Understand how the characteristics of a data-set affects performance of a model.

1.5 Assumptions & Limitations

Though an efficient model has been proposed to predict lung cancer based on given feature, it has limitations on which further studies should be done:

- There are more powerful machine learning algorithm, i.e Gradient Boosting (including XGBoost, LightGBM, and CatBoost) which can be used. As we aim to create a project based on what we learn in classroom, we did not include any highly efficient machine learning algorithm.
- Data found on kaggle was used for assessing the model performance.

- Three of the datasets are small-sized. The only large dataset has very poor feature correlation which failed to assess model efficiency.
- No comparison among different combination of models has been analyzed so can't be declared it as the optimal way.
- Only accuracy is used as performance measure to evaluate performance of classifiers.
- Only grid search is used as hyper-parameter tuning for finding the optimal number of neighbors in KNN algorithm, but each of the algorithm can have optimal parameters by using hyper-parameter tuning.
- The data-set being small put a remarkable challenge for model being over-fitting.

1.6 Research Outline

Rest of the report is structured as follows: In **Chapter II** a literature study on related work is given. **Chapter III** introduces the machine learning model describing each of its steps. **Chapter IV** discusses about result on model performance for each data-set, including the performance of ensemble learning for hard-voting and soft-voting. Lastly in **Chapter V** future work and conclusion is mentioned.

CHAPTER II

Literature Review

2.1 Performance Analysis using Machine Learning techniques

2.1.1 Logistic Regression

Logistic regression will predict the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it will give the probabilistic values which lie between 0 and 1.

2.1.2 Decision Tree

The internal nodes of the Decision tree will represent the features of a dataset, branches represents the decision rules and each leaf node represents the outcome. It will simply ask a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

2.1.3 Random Forest

Random Forest will contain a number of decision trees on various subsets of the given dataset and will take the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest will take the prediction from each tree and based on the majority votes of predictions, and it will predict the final output.

2.1.4 SUpport Vector Classifier

SVM will choose the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors. It will handle outliers.

2.1.5 K-Nearest Neighbour

The KNN algorithm will assume the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.

2.1.6 Gaussian Naive Bayes

It will assume the features of a data point are independent of each other. As It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

2.1.7 High Performance Filter

High performance filter is a term that can refer to different techniques for feature selection in machine learning. Feature selection is the process of choosing a subset of relevant features from the original data that can improve the performance and efficiency of machine learning models. High performance filter methods are those that use some criteria or measure to rank the features according to their importance or relevance, without involving any learning algorithm. Some examples of high performance filter methods which will be used in the proposed model are Random forest approach, Double input symmetrical relevance filter (DISR), Joint impurity filter (JIM). High performance filters will enhance the generalization ability of machine learning models of the proposed system.

2.1.8 Ensemble Learning

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Basic idea is to learn a set of classifiers (experts) and to allow them to vote.[ensem] Voting Classifier supports two types of votings.

Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.[3]

Soft Voting: In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.[3]

2.1.9 Cross-validation

K-Fold cross validation is a popular technique used in machine learning for model evaluation and selection. It involves dividing a dataset into K subsets of equal size, called folds. The algorithm then trains and evaluates the model K times, each time using a different fold as the validation set and the remaining K-1 folds as the training set.During each iteration of K-Fold cross validation, the model is trained on K-1 folds and evaluated on the remaining fold. The performance metrics are then averaged over all K iterations to obtain an estimate of the model's overall performance.[4]

2.1.10 Stratified Cross-Validation

Stratified K-Fold cross-validation is a modification of the standard K-Fold cross-validation technique that is commonly used in machine learning when working with imbalanced datasets. The goal of Stratified K-Fold cross-validation is to ensure that each fold is representative of the overall dataset in terms of the class distribution. In standard K-Fold cross-validation, the data is split into K folds, and each fold is used as the validation set in turn. However, if the dataset has an imbalanced class distribution, this can lead to some of the folds having significantly fewer samples from the minority class, which can result in biased performance estimates. To address this issue, Stratified K-Fold cross-validation ensures that each fold has a similar proportion of samples from each class. It works by first dividing the dataset into K folds, as in standard K-Fold cross-validation. Then, for each fold, the algorithm ensures that the proportion of samples from each class is roughly the same as the proportion in the full dataset. This ensures that the model is evaluated on a representative sample of the data, regardless of the class distribution.[4]

2.2 Existting works on dataset

2.2.1 Existing works on Dataset 1

The first dataset is 'survey lung cancer.csv'[2], this is the most upvoted dataset having 309 rows in total. There are 80 works on it. Among them, some remarkable works includes the following: Anju Sukumaran, at [5] analyzed the data and predicted lung cancer using a support vector classifier with 96% accuracy. Abocadobaby, at [6], predicted lung cancer using adaboost classifier and GridSearchCV with 97% accuracy. Shirshendu Dey, at [7] implemented a machine learning model that uses decision tree classifiers to predict lung cancer using. Adaikkappan, at [8] uses a random forest algorithm to predict lung cancer. Karen1116, at [9], used various machine learning algorithms, i.e. logistic regression, random forest classifier, K neighbor classifier, decision tree classifier, gradient boost classifier, XGB classifier with accuracy 64%, 99.07%, 93.51%, 95.37%, 96.296%, 99.07% respectively. In this project he showed comparison among six machine learning model performances. Sandra Grace Nelson ,at [4], trains ten machine learning models, each model is trained and tested multiple times. Then the average their accuracy is taken as the accuracy for that particular model. She used Logistic regression (92.88%), Decision tree (92 %), KNN(91.84%), Gaussian naive bayes(88.07%), Multinomial naive bayes (75.72%), Support Vector Classifier models (94.76%), Random forest model (94.56%), XGBoost model (94.57%), Multilayer perceptron model (93.92%), Gradient boost models (94.76%).

2.2.2 Existing works on Dataset 2

The next dataset is 'survey lung cancer.csv'[9]. It is actually a bigger version of the previous dataset having 55400 rows. Though it has been upvoted 22 times, there is no publicly available work on it.

2.2.3 Existing works on Dataset 3

The third dataset is 'lung cancer.csv'[3], is a subset of data available from the US National Lung Screening Trial (NLST). This dataset mainly concentrates on the risk of lung cancer because of smoking. There is 10 publicly available work on it, among these, Tetsuya Sasaki at [10], has run Logistic regression, KNN, Random Forest algorithm on this dataset separately. Among these the maximum testing accuracy was 59%. Later he performed principal component analysis which results in a little bit better accuracy of 60%.

2.2.4 Existing works on Dataset 4

The fourth dataset is 'cancer patients data sets.csv'[11]. There are 1000 rows and 26 columns. Sonu Sain at [12] run random forest algorithm resulting in 100% accuracy, whereas, Anagha K P at [13] did not scale the data and run LR algorithm having 90% accuracy.

2.3 Research Gap

Among all these works, we saw different algorithms having different strengths of predicting correctly. None of these works combine the strengths which can make the model more generalized and robust. Hence, the hybrid machine learning model can predict cancer with better accuracy for data with different patterns. Our project addresses this issue and attempts to obtain the aim of implementing a novel model using machine learning, high performance filter and ensemble learning. Also, we found some works used some of the features like scaling, grid search, principal component analysis, cross-validation. Applying this features while needed can improve the scope of robustness.

CHAPTER III

Methodology

We run the model on four different datasets. The datasets with their descriptions are given below. Later, the steps in the model will be discussed.

3.1 Dataset

3.1.1 Dataset 1

The dataset 1, 'survey lung cancer.csv'[2], has 309 rows and 16 columns.This dataset is mainly focus on how various symptoms can be related to probability of lung cancer.

This dataset has a data imbalance issue regarding the target column, 'Lung Cancer' which has been handled using adaptive synthetic sampling.

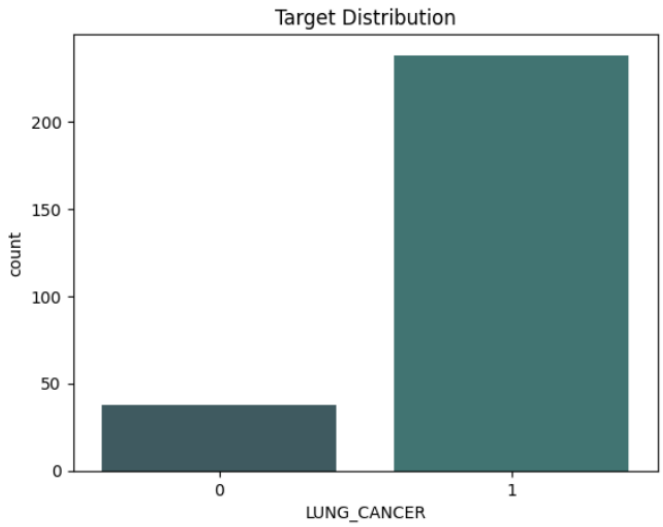

Figure 3.1: Imbalance in dataset

Attribute	Possible Values
Gender	Female, Male
Age	**Age of Patient**
Smoking	**YES=2 , NO=1.**
Yellow fingers	**YES=2 , NO=1.**
Anxiety	YES=2 , NO=1.
Peer_pressure	YES=2 , NO=1.
Chronic Diseases	YES=2 , NO=1.
Fatigue	YES=2 , NO=1.
Allergy	YES=2 , NO=1.
Wheezing	YES=2 , NO=1.
Alcohol	YES=2 , NO=1.
Coughing	YES=2 , NO=1.
Shortness of Breath	YES=2 , NO=1.
Swallowing Difficulty	YES=2 , NO=1.
Chest pain	YES=2 , NO=1.
Lung Cancer	YES=2 , NO=1.

Table 3.1: Attributes of dataset 1

We found that The correlation matrix shows that ANXIETY and YELLOW_FINGERS are correlated more than 50%. Therefore, a new feature has been created combining them.

3.1.2 Dataset 2

The dataset 2[14] has 55394 rows and 16 columns. From figure 3.3 shows this dataset having 1.62% duplicates where 0.017% has 2 duplicates in total, 1.6% has 1 duplicate . Its attributes are the same as the previous database with a balanced target column. The feature of this dataset is not correlated enough to derive pattern. This can affect model performance as is seen in figure.

3.1.3 Dataset 3

The dataset 3[3] is focused on the level of lung cancer (cancer stage) among smoker based on seven attributes as following: Among these 53427 rows of this dataset, 51395 rows have 'nan' value at the target column as seen in figure 3.7. Filling such a vast number of target variables may not represent the actual underlying data. Therefore, we removed the rows resulting in a dataset of 2032 rows,7 columns.

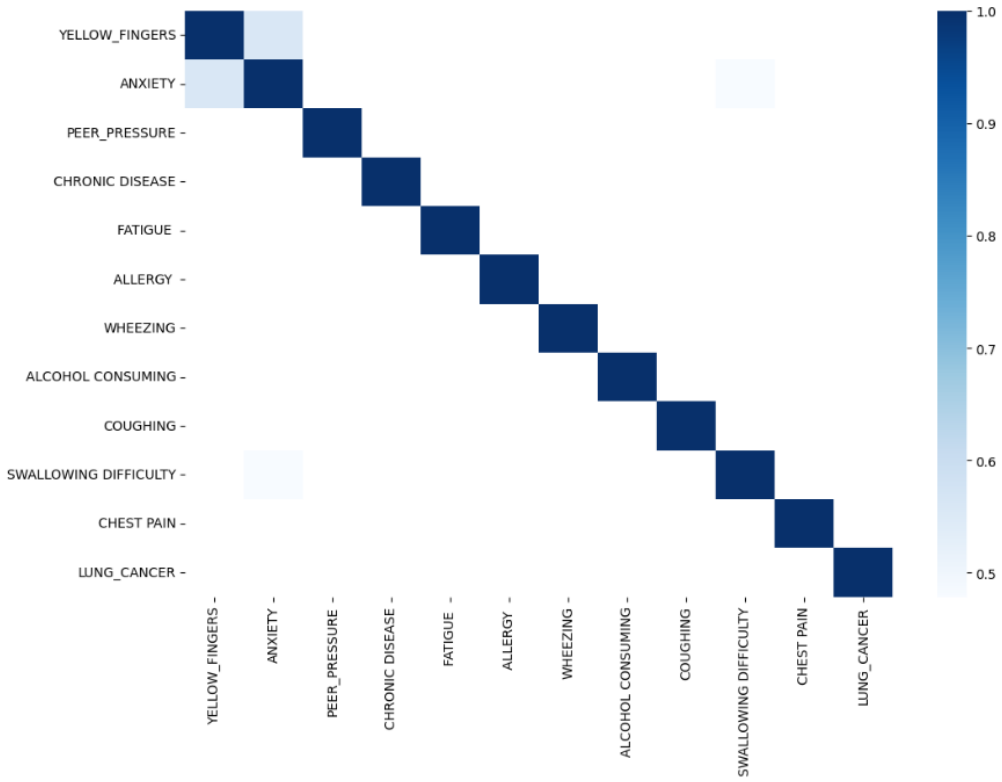

Figure 3.2: Features in the dataset that are correlated more than 50%.

3.1.4 Dataset 4

The dataset 4[11] contains information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy (the level of dust allergy of the patient), occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss ,shortness of breath ,wheezing ,swallowing difficulty ,clubbing of finger nails and snoring. Except for age, each of these features are categorical. Though the dataset does not have a data imbalance issue, 848 rows out of 1000 rows are duplicates.

3.2 System Model

After the data is pre-processed, data is fed to each of the six machine learning algorithm. To observe the difference and for efficient assessment of model performance all of these six model is run with K-Fold cross validation for each of the four datasets. Further, stratified cross-validation is used to have more generalized performance.

Attributes	Values
Pid	Anonymous identifier of a person
Age	Age of the person
Gender	Male / Female
Race	Race of the person (White, Black, Others)
Smoker	Former smoker/ Current smoker
Days_of_cancer	Number of days passed since the trial when the cancer was first observed
Stage_of_cancer	The stage of cancer when the cancer was first observed. Four stages in total.

Table 3.2: Attributes of dataset 3

Attribute name	Description
Age	The age of the patient. (Numeric)
Gender	The gender of the patient. (Categorical)
Air Pollution	The level of air pollution exposure of the patient. (Categorical)
Alcohol use	The level of alcohol use of the patient. (Categorical)
Dust Allergy	The level of dust allergy of the patient. (Categorical)
OccuPational Hazards	The level of occupational hazards of the patient. (Categorical)
Genetic Risk	The level of genetic risk of the patient. (Categorical)
chronic Lung Disease	The level of chronic lung disease of the patient. (Categorical)
Balanced Diet	The level of balanced diet of the patient. (Categorical)
Obesity	The level of obesity of the patient. (Categorical)
Smoking	The level of smoking of the patient. (Categorical)
Passive Smoker	The level of passive smoker of the patient. (Categorical)
Chest Pain	The level of chest pain of the patient. (Categorical)
Coughing of Blood	The level of coughing of blood of the patient. (Categorical)
Fatigue	The level of fatigue of the patient. (Categorical)
Weight Loss	The level of weight loss of the patient. (Categorical)
Shortness of Breath	The level of shortness of breath of the patient. (Categorical)
Wheezing	The level of wheezing of the patient. (Categorical)
Swallowing Difficulty	The level of swallowing difficulty of the patient. (Categorical)
Clubbing of Finger Nails	The level of clubbing of finger nails of the patient. (Categorical)

Table 3.3: Attributes of dataset 4

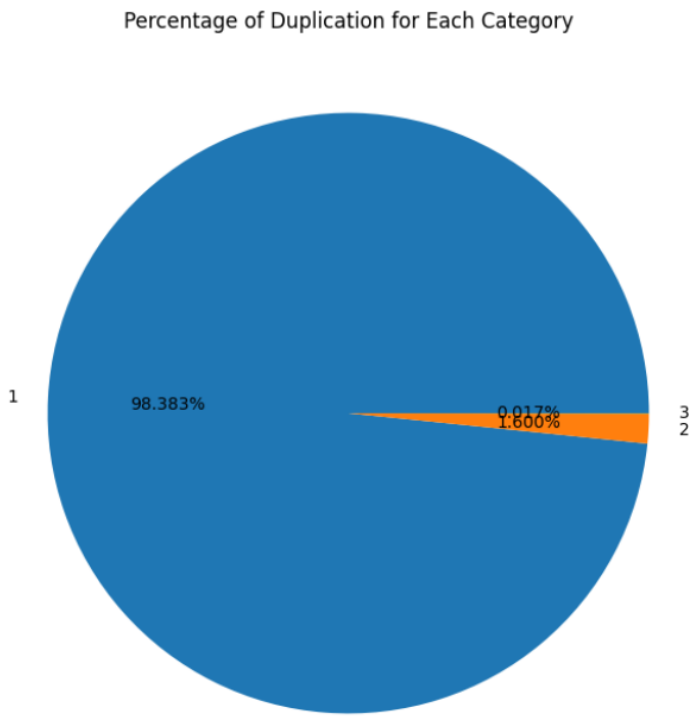

Figure 3.3: Percentages of duplicates in dataset 2.

Then, the top best performing machine learning models are used as estimators in the next phase for ensemble learning. Ensemble learning is also performed without cross-validation and with cross-validation.

This model introduces a comparison strategy among various classification methods of machine learning technology and how ensemble learning can be used to evaluate the overall performance.The following figure 3.1 illustrates the flow of each step of the model.

3.2.1 Data Preprocessing

- **Checking null values:** Dataset 3 contains 96.19% null values in the target column. Removing those results in 2032 rows which is still a considerable number. Filling up those nan with mean value of 2032 may not represent, may even be biased. Because of these reasons, it seems more reasonable to remove this instances.

- **Handling duplicates:** Small number of duplicates can be part of actual data distribution. But a large number of duplicates can corrupt the actual underlying distribution. The percentage of duplicates in Dataset 2 is negligible whereas

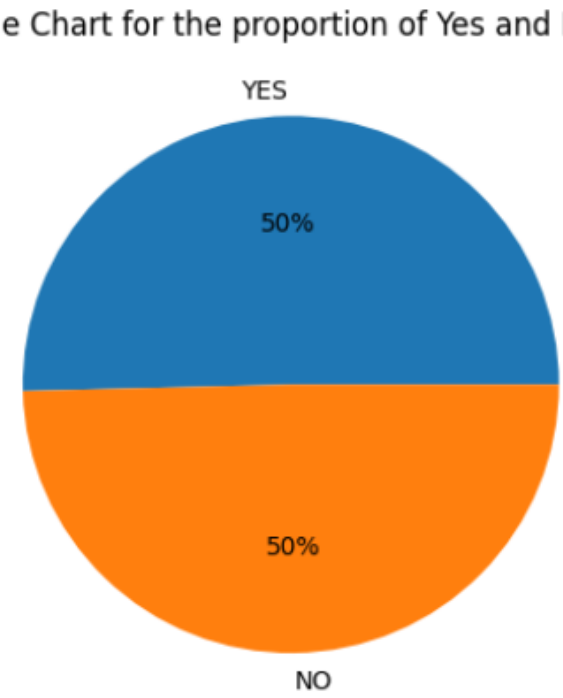

Figure 3.4: Balanced dataset

84.8% of dataset 4 is duplicated data. How the duplicated values are spread can affect the biases of the dataset.

- **Data scaling:** Before applying machine learning algorithm, it is important to scale the data points. Because when features are on different scales, it can lead to problems in the learning process, particularly for algorithms that rely on distances, gradients, or optimization, i.e., k-nearest neighbor, logistic regression, principal component analysis, support vector machine etc.

- **Feature engineering:** For having better model accuracy, feature engineering has been applied by combining two highly correlated feature to create a new feature, as needed.

3.2.2 Performance analysis of machine learning model

After preprocessing, dataset is fed to each of the six machine learning algorithms, Logistic Regression (LR), Decision Tree classifier (DT), Random Forest classifier (RF), K-Nearest Neighbors classifier (KNN), Gaussian Naive Bayes (GNB), Support Vector Classifier (SVC). To observe the difference and avoid the algorithms are run with a k-fold cross validation process. The performance metrics are then averaged

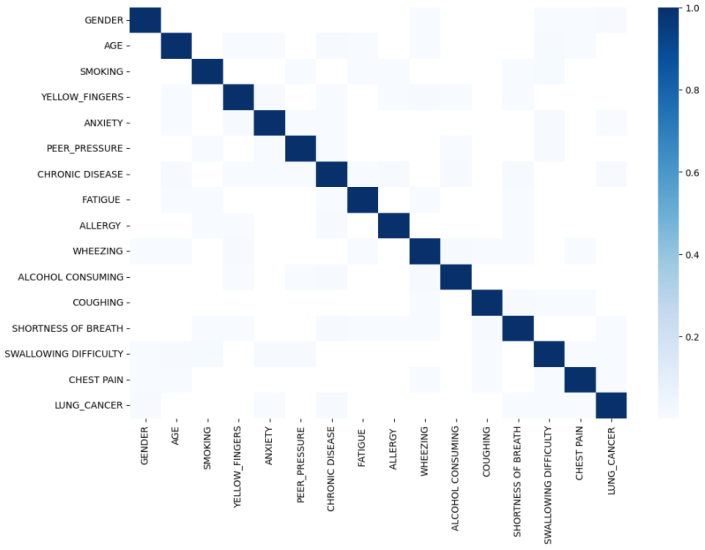

Figure 3.5: Features that are correlated more than 0.1%.

over all K iterations to obtain an estimate of the model's overall performance. K-Fold cross validation is a robust method for model evaluation because it uses all the available data for training and testing. It also helps to reduce the risk of overfitting and provides a more accurate estimate of the model's performance than using a single training-test split.

3.2.3 High performance filter

Based on the performance, top three machine learning algorithms are selected which have the highest performance. The selected algorithms are combined for ensemble learning in the next phase.

3.2.4 Ensemble Learning

We observe the performance of ensemble learning-

- Without cross-validation
- With cross-validation
- After performing principal component analysis (only for dataset 3)

Based on these observations we conclude on model performance regarding the nature of the dataset.

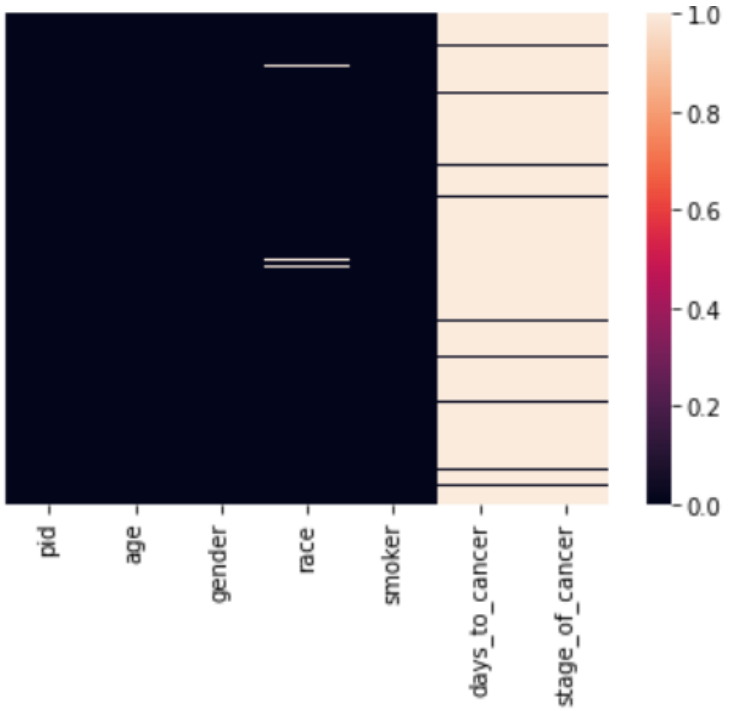

Figure 3.6: Null values in dataset 3.

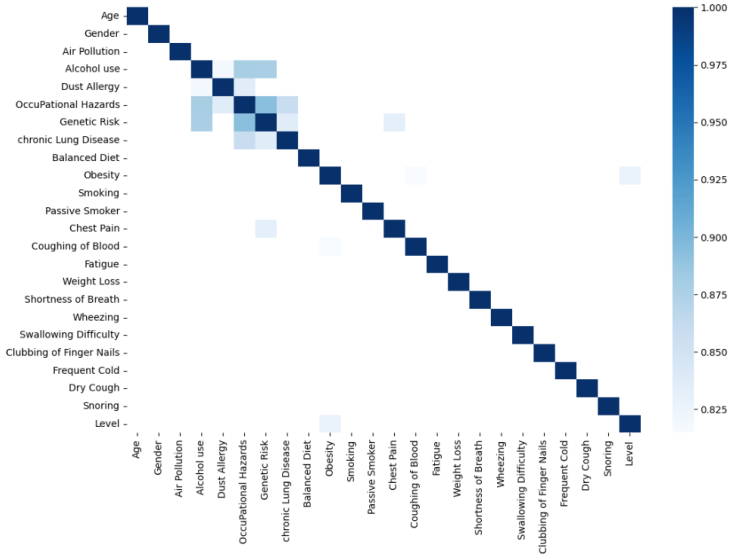

Figure 3.7: Features in dataset 4 that are correlated more than 80%.

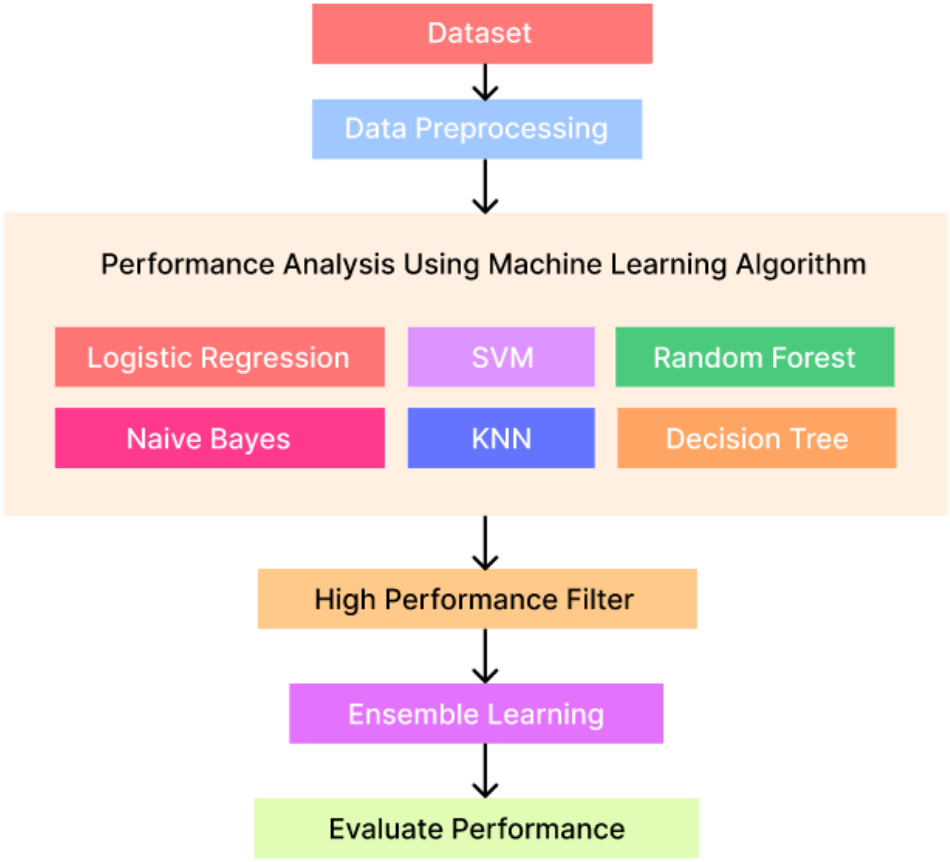

Figure 3.8: Structure of the Model

CHAPTER IV

Result and Discussion

4.1 Result per Dataset

4.1.1 Dataset 1

Summary of performance of the machine learning algorithms: We can observe from tabel 4.1 that as cross validation is being used, accuracy started to fall. This can be because of data variability. Cross-validation provides a more realistic estimate of how the model will perform on unseen data. **Result of ensemble learning:** We can see from tabel 4.2, as ensemble learning is being used, better accuracy can be achieved. But with cross-validation the accuracy is better than most of the individual algorithm but it still less than the accuracy of ensemble learning without cross-validation. One probable reason for that is, both cross-validation and ensemble learning is efficient when the dataset is larger. Dataset 1 being a smaller dataset cannot illustrate the efficiency.

Algorithm	Without Cross-validation	With Cross-validation	With stratified Cross-validation
LR	**97%**	**93%**	**92.88%**
DT	**94%**	**95%**	**92.3%**
KNN	**96%**	**93%**	**91.8%**
GNB	**92%**	**88.5%**	**88.7%**
SVC	**98%**	**94.97%**	**94.76%**
RF	**98%**	**95.4%**	**94.35%**

Table 4.1: Accuracy of each machine learning algorithms on Dataset 1.

Voting Type	Without cross-validation (Accuracy)	With cross-validation (Accuracy)
Hard-voting	**97.5%**	**94.537%**
Soft-voting	**99.16%**	**94.537%**

Table 4.2: Accuracy of ensemble learning algorithms on Dataset 1.

Algorithm	Without Cross-validation	With Cross-validation
LR	**49%**	**49.98%**
DT	**50%**	**50.12%**
KNN	**50%**	**50.17%**
GNB	**50%**	**49.92%**
SVC	**51%**	**50.39%**
RF	**51%**	**50.46%**

Table 4.3: Accuracy of each machine learning algorithms on Dataset 2.

4.1.2 Dataset 2

Summary of performance of the machine learning algorithms: From table 4.3 it is clear that the top three high performance machine learning model: K-nearest neighbor, support vector machine, random forest classifier. Ensemble learning will be performed using these three classifier. **Result of ensemble learning:** Clearly, tabel 4.4 shows that both of these voting gives almost similar result which is higher than most of the accuracy found from each individual machine learning while cross-validation was applied.

4.1.3 Dataset 3

Summary of performance of the machine learning algorithms: Clearly, from table 4.5 we can see, when cross validation is being used, accuracy started to fall. This can be because of data variability. Also, using stratified cross validation the accuracy increased. Stratified cross-validation helps prevent issues related to class imbalances and can improve the model's generalization. The top performing machine

Voting-type	Hard-voting	Soft-voting
Accuracy	**50.494%**	**50.422%**

Table 4.4: Accuracy of ensemble learning algorithms on Dataset 2.

Algorithm	Without Cross-validation	With Cross-validation	With stratified Cross-validation
LR	**60%**	**57.4%**	**57.28%**
DT	**50%**	**52.85%**	**52.66%**
KNN	**51%**	**53.08%**	**55.07%**
GNB	**60%**	**56.34%**	**56.5%**
SVC	**60%**	**59.06%**	**59.3%**
RF	**54%**	**54.23%**	**55.26%**

Table 4.5: Accuracy of each machine learning algorithms on Dataset 3.

Voting Type	Without cross-validation (Accuracy)	With cross-validation (Accuracy)
Hard-voting	**54.55%**	**57.18%**
Soft-voting	**59%**	**56.29%**

Table 4.6: Accuracy of ensemble learning algorithms on Dataset 3.

learning models: Support vector classifier, Logistic regression, Gaussian Naive Bayes are selected for ensemble learning.

Result of ensemble learning: (Using PCA) From table 4.6 we can say that though soft-voting gives better accuracy while cross-validation is not used, hard-voting seems to work better while cross-validation is used.

4.1.4 Dataset 4

Summary of performance of the machine learning algorithms: From table 4.7 we can see that the top three highest performing algorithms Linear regression, Decision Tree classifier, Random Forest Classifier are selected to act as estimators in ensemble learning at next phase.

Algorithm	Without Cross-validation	With Stratified Cross-validation	With stratified Cross-validation
LR	**99.27%**	**100%**	**57.28%**
DT	**100%**	**100%**	**52.66%**
KNN	**88.69%**	**90.8%**	**55.07%**
GNB	**81%**	**83.89%**	**56.5%**
SVC	**100%**	**90.5%**	**59.3%**
RF	**100%**	**100%**	**55.26%**

Table 4.7: Accuracy of each machine learning algorithms on Dataset 4.

Voting Type	Without cross-validation (Accuracy)	With Stratified cross-validation (Accuracy)
Hard-voting	**100%**	**100%**
Soft-voting	**100%**	**100%**

Table 4.8: Accuracy of ensemble learning algorithms on Dataset 4.

Result of ensemble learning: Clearly, from table 4.8 we can say that the model has been overfitted. One most probable reason can be its smaller size with 848 duplicates within 1000 samples.

4.1.5 Observations

- Cross-validation has been used to have stable results on performance. This tends to reduce individual accuracy.
- Dataset 3 was hard to interpret, using principal component analysis helped increasing accuracy.
- Both ensemble learning and cross-validation is efficient while the dataset is large. As 3 of our datasets consist of 309 rows, 1000 rows and 2024 rows, the effect of using these technique cannot be seen as remarkable.
- The dataset 2 has almost 55000 rows, but poor correlations among features of the dataset badly affected the learning capability of machine learning algorithms.

CHAPTER V

Future Work & Conclusion

5.1 Future Work

This project does not include algorithms like gradient boosting,cat boosting which can increase the ability of predictive task. Also, the dataset 2 has poor correlated feature which needs to be handle. This can be focused on future working on this project. Hyper-parameter tuning can be applied to each of the six algorithm to have optimal parameter most effective for a specific dataset. This can make this model more robust and accurate.

5.2 Conclusion

Though we observed the increase in performance of the model after using ensemble learning in some cases, we have also seen overfitting. Throughout the process, feature engineering, the model utilizes grid search for optimal value of k-nearest neighbor, k–fold cross validation (also, in some case stratified cross-validation) both for measuring machine algorithm performance as well as for ensemble performance for model selection. Thus, generalization along with a high level of robustness have been gained. Only model can not derive best accuracy, dataset nature plays a great role here. Small dataset tends to be biased, can result in overfitting. As machine learning learns from data, larger data is desired.

References

[1] D. R. Baldwin, ""prediction of risk of lung cancer in populations and in pulmonary nodules: Significant progress to drive changes in paradigms," *Lung Cancer*, 2015.

[2] A. Sofyan, "Survey lung cancer." `https://www.kaggle.com/datasets/ajisofyan/survey-lung-cancer`. Accessed on 2023 Aug. 31.

[3] RADDAR, "Smoking related lung cancers." `https://www.kaggle.com/datasets/raddar/smoking-related-lung-cancers/cod`. Accessed on 2023 Oct. 25.

[4] Sandra, "Lung cancer prediction." `https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction`. Accessed on 2023 Aug. 31.

[5] A. Sukumaran, "Lung cancer prediction eda." `https://www.kaggle.com/code/abocadobaby/lungcancer-prediction-f1-score-97`. Accessed on 2023 Aug. 31.

[6] abocadobaby, "Lungcancer prediction (f1-score : 97)." `https://www.kaggle.com/code/abocadobaby/lungcancer-prediction-f1-score-97`. Accessed on 2023 Aug. 31.

[7] S. Dey, "Lung cancer eda and decision tree classifier." `https://www.kaggle.com/code/shirshendudey99/lung-cancer-eda-and-decision-tree-classifier`. Accessed on 2023 Aug. 31.

[8] A. A, "Lung cancer predictions by using random forest." `https://www.kaggle.com/code/adaikkappana/lung-cancer-predictions-by-using-random-forest`. Accessed on 2023 Aug. 31.

[9] Karen1116, "Lung cancer predictions." https://www.kaggle.com/code/karen1116/lung-cancer-predictions. Accessed on 2023 Aug. 31.

[10] T. SASAKI, "Smoker lung cancer stage detection model." https://www.kaggle.com/code/sasakitetsuya/smoker-lung-cancer-stage-detection-model. Accessed on 2023 Oct. 25.

[11] T. DEVASTATOR, "Lung cancer prediction." https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data. Accessed on 2023 Oct. 25.

[12] S. SAINI, "Lung cancer prediction." https://www.kaggle.com/code/raman209/lung-cancer-prediction. Accessed on 2023 Oct. 25.

[13] A. K. P, "Lung cancer prediction(logistic regression model)." https://www.kaggle.com/code/anaghakp/lung-cancer-prediction-logistic-regression-model. Accessed on 2023 Oct. 25.

[14] M. A. BHAT, "Lung cancer." https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Accessed on 2023 Oct. 25.