Cancer Prediction Using Random Forest with Lifestyle Factors

SHAM ASWIN KARMUKILAN D K SANTHOSH KUMAR

Abstract

With the help of various machine learning algorithms and lifestyle variables, this research attempted to predict cancer. In order to evaluate the classification outcomes of the Random Forest, Logistic Regression, Support Vector Machines, and KNN algorithms, we have chosen our dataset from the Kaggle repository. Correctly categorized instances, incorrectly classified instances, F-Measure, Precision, Accuracy, and Recall were among the classification parameters. According to our findings, Random Forest did better than the other algorithms for large datasets and the same number of attributes. While Random Forest achieved 100% precision, the other algorithms were less accurate. This study is important because accurate cancer classification and early detection can greatly lower mortality rates.

Keywords: Cancer, Machine learning algorithms, Random Forest, Logistic Regression, Support Vector Machines, KNN.

1. Introduction

The Random Forest algorithm is a popular machine learning technique that has found applications in various fields such as computer vision, text classification, and medical diagnosis. In this article, we will explore the Random Forest algorithm and its use in predicting cancer with lifestyle factors.The Decision Tree algorithm is a popular machine learning algorithm also used for classification and regression tasks. Decision Trees are constructed by recursively partitioning the input space into subsets, based on the values of one or more input features. The final result is a tree-like structure where each leaf node corresponds to a classification label Random Forest is an extension of the Decision Tree algorithm that overcomes some of these limitations. Random Forests use a collection of Decision Trees, where each tree is trained on a subset of the data and a random subset of the features. The final prediction is made by aggregating the predictions of all the trees.

The benefit of Random Forest includes:

- Overcoming the issue of overfitting;
- Categorical, and binary data, making it suitable for high and Less sensitivity to aberrant data in training data.
- Setting parameters is simple, so there is no need to prune the plants.
- Automatically generated variable accuracy and significance

In this article, we concentrate on the classificationperformance of the Random Forest for the prediction. The objective of this comparison is creating a base-line, which will be useful forthe classification scenarios of prediction of the cancer. It will also help in the selection of appropriate model.

The remaining portions of the paper are structured as follows: Classification methods, such as the Random Forest, are described in Section 2. The Section 3 description of the experimental setting and the datasets used. The findings and conclusion are shown in Section 4.

2. Classification Methods

2.1 Random Forest

In machine learning, the widely used ensemble learning method Random Forest is used for both classification and regression problems. A final prediction is made after the construction and combining of numerous trees, which is an extension of the decision tree algorithm. Each Decision Tree in a Random Forest is constructed using a randomly chosen portion of the training data and a randomly chosen subset of the features. Every tree in the forest receives autonomous training before making a prediction. By combining the predictions made by each tree in the forest, the final prediction is achieved. Averaging or majority polling are two techniques that can be used for the aggregation. Figure 1 describes the workflow of Random Forest algorithm.

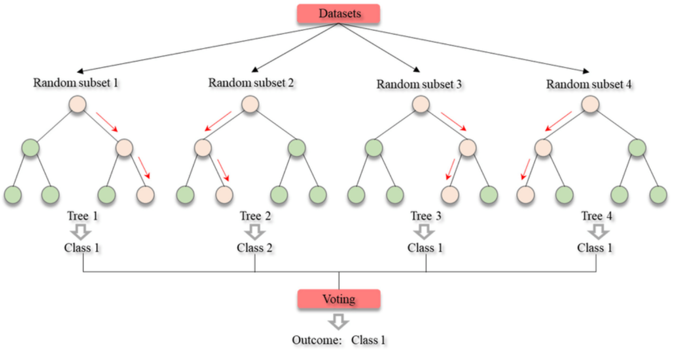


Figure 1: Working of Random Forest [15]

Decision Trees are built from pre-classified data and focus on finding the best split features to classify the data, Random Forest is an ensemble learning method that builds multiple Decision Trees using randomized subsets of the data and features to improve the model's accuracy and robustness. Random Forest is an effective method for dealing with missing data, noisy data, and outliers. The process of building multiple trees and aggregating their predictions helps to smooth out the effects of noise and outliers, resulting in a more stable and accurate model.

3. Experimental Analysis

In this section, we concentrate on the classification performance of the various algorithms. The objective of this comparison is creating a base-line, which will be useful for the classification scenarios. It will also help in the selection of appropriate model.

3.1 Data Sets

The dataset contains 25 columns out of which the Patient ID column is left out as it is not relevant to the experiment. The ‘Level’ column further is taken as our target variable. All of the other instances does not contains any erroneous or missing data, and thus all 1000 entries are included in the training data. The final dataset contains 1000 entries included where there are no null or missing values. The dataset is split into training and testing sets using the train_test_split() function from the scikit-learn library. The testing set size is set to 40% of the total dataset, and the random_state parameter is set to 0 for reproducibility.

The training set is further split into development and validation sets using the same train_test_split() function. The validation set size is set to 30% of the training set, and again, the random_state parameter is set to 0.

Table : Attributes in the dataset (Lifestyle Factors)

AGE	Balanced Diet	Shortness of Breadth
Gender	Obesity	Wheezing
Air pollution	Smoking	Swallowing Difficulty
Alcohol Usage	Passive Smoking	Clubbing of Finger Nails
Dust Allergy	Chest Pain	Frequent Cough
Occupational Hazards	Coughing	Dry Cough
Genetic Disorders	Fatigue	Snoring
Chronic Disease	Weight Loss	

Feature Selection

The feature correlation matrix shows the correlation coefficients between pairs of features in the dataset. The coefficient ranges from -1 to 1, where a value of 1 represents a strong positive correlation, 0 represents no correlation, and -1 represents a strong negative correlation.Looking at the matrix, we can see that there are some strong positive correlations between certain features, such as OccuPational Hazards and Alcohol use (0.88), Genetic Risk and OccuPational Hazards (0.89), and Obesity and Coughing of Blood (0.81). There are also some moderate positive correlations between features, such as Air Pollution and Alcohol use (0.75), Dust Allergy and OccuPational Hazards (0.84), and Chest Pain and Genetic Risk (0.83).On the other hand, there are some features that have no correlation with each other, such as Gender and Frequent Cold (-0.0).

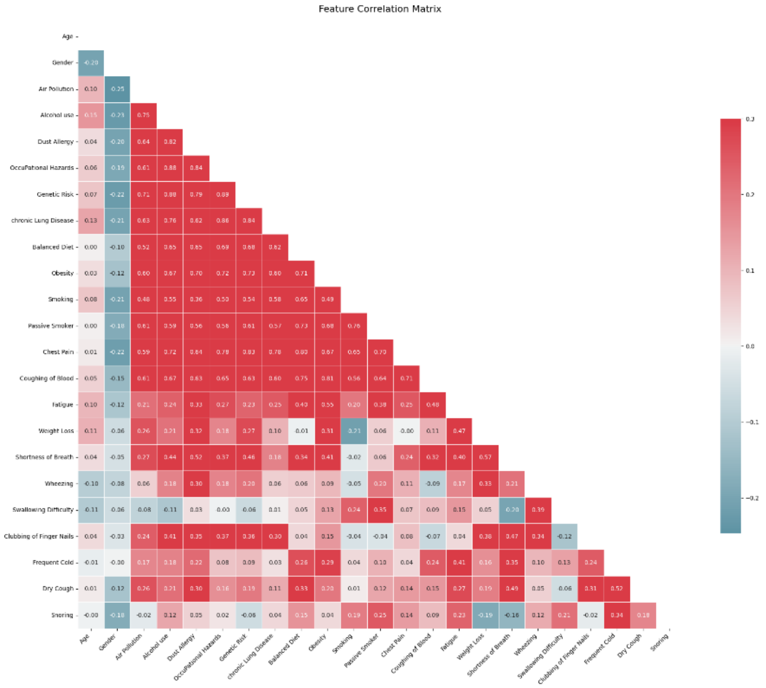


Figure 2: Feature Correleation Matrix

Model Development

The Random Forest algorithm is chosen for model training based on the findings of the other algorithms, such as Logistic Regression, Support Vector Machines, and KNN. Random Forest has an accuracy of 1.0 in both the validation and test sets, showing that the model can perfectly classify the data. The classification report also indicates that the precision, recall, and f1-score for all classes are all 1.0, supporting this. On the test set, the model has an accuracy of 0.97 for Logistic Regression, which is still quite excellent. According to the classification report, the precision, recall, and f1-score for all classes are also quite good. On the test set, the model has an accuracy of 0.97 for KNN, which is also good. According to the classification report, the precision, recall, and f1-score for all classes are also quite good. On the test set, SVM has an accuracy of 0.93, which is lower than the other models. According to the classification report, the precision, recall, and f1-score for the High and Medium classes are very excellent, but the recall for the Low class is marginally lower.

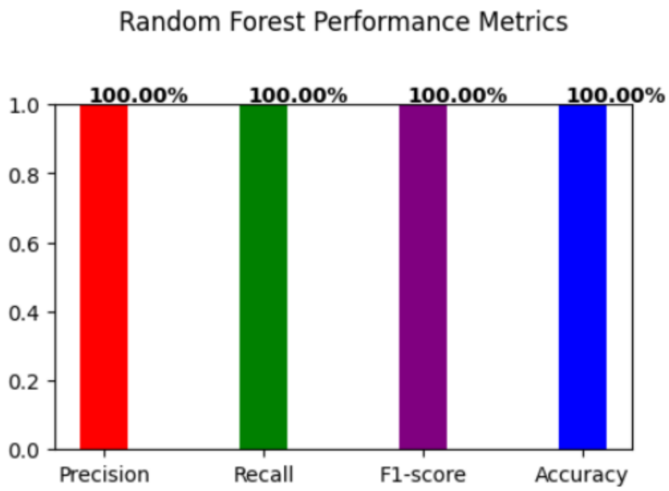


Figure 3 : Random Forest Results

Figure [3], [4], [5], [6] has graph plots with the x-axis indicates performance metrics including precision, recall, F1-score, accuracy and the y-axis represents the values of the metrices. Even though the graph of SVM plots are 100 percent in accuracy the findings of precision and f-1 score are less as compared to that of the Random Forest.

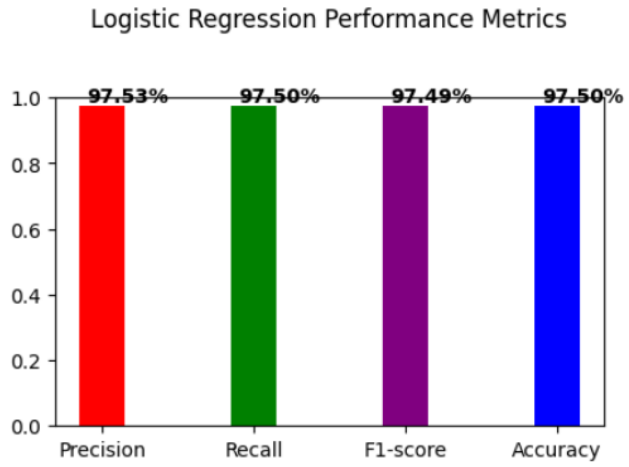


Figure 4 : Logistic Regression Results

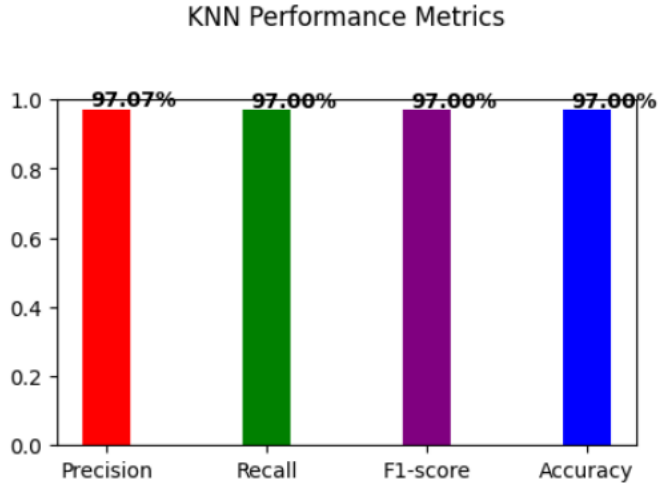


Figure 5 : KNN Results

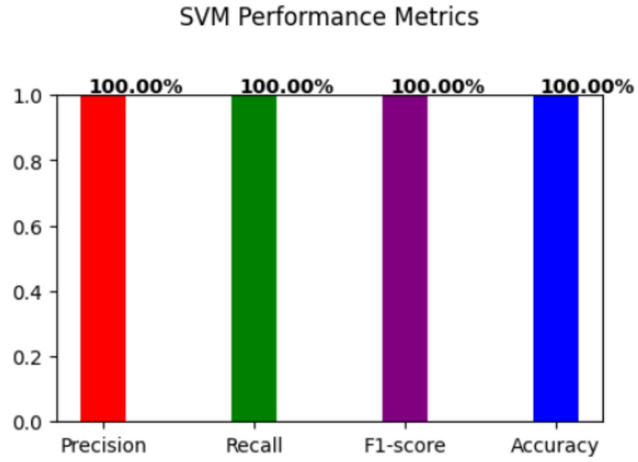


Figure 6 : Support Vector Machine Results

Model Evaluation

The four models are trained and evaluated on both the initial and minimal datasets. As accuracy is not a sufficient metric in the medical field, the models are compared based on a multitude of measures. The particular metrics used are detailed below.

Basic definitions

- True Positive (TP): The count of Cancer classified as Cancer by the model.
- True Negative (TN): The count of Low Cancer classified as low by the model.
- False Positive (FP): The count of Low Cancer classified as High by the model.
- False Negative (FN): The count of High classified as Low by the model.

Metrics used.

- Accuracy: The percentage of cancer which was correctly predicted.
- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- F1 Score: (Precision * Recall) / (Precision + Recall)

All four models are trained and evaluated using these metrics datasets. Their performances are collated and analyzed in the Results section.

4. Results and Discussion

The below Confusion matrix Figure [7] and the table presents the performance metrics of four machine learning models- Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) on the dataset.

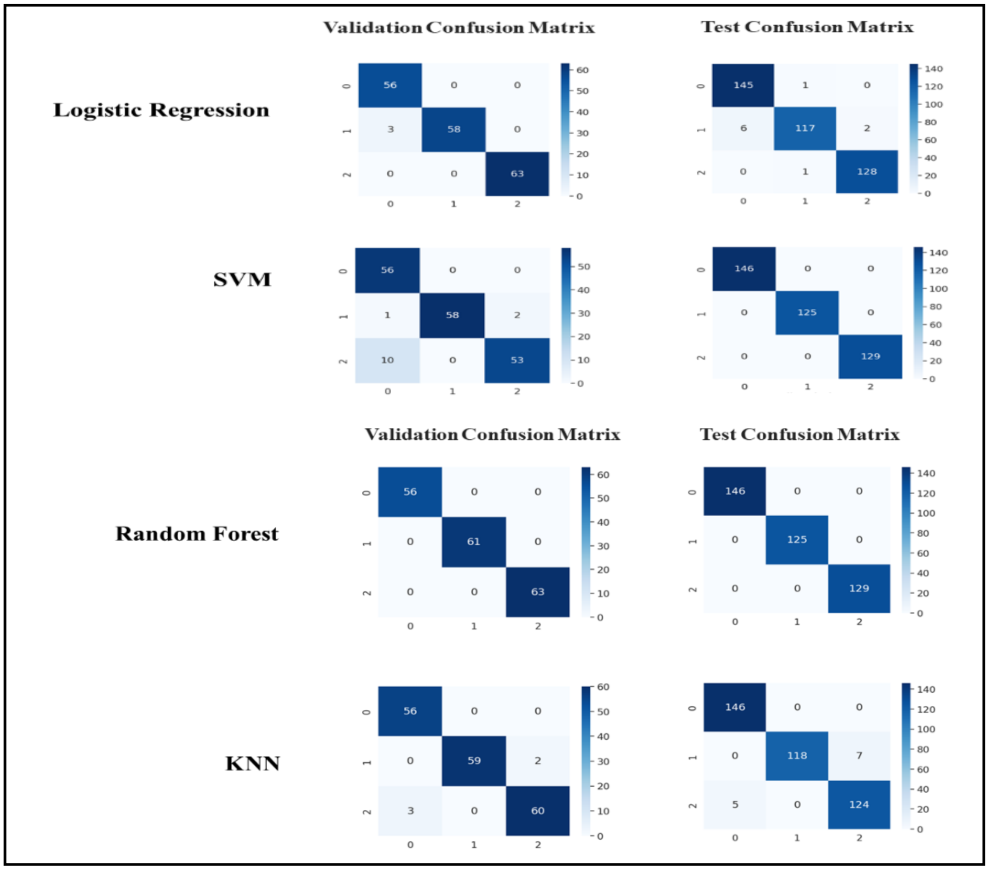
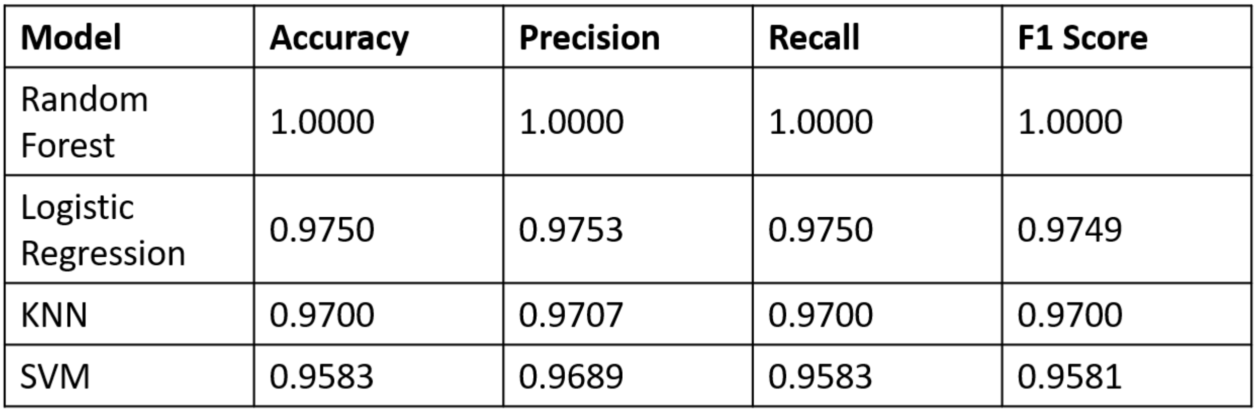


Figure 7 : Confusion Matrix

Table : Model Comparison on the dataset



Model	Accuracy	Precision	Recall	F1 Score
Random Forest	1.0000	1.0000	1.0000	1.0000
Logistic Regression	0.9750	0.9753	0.9750	0.9749
KNN	0.9700	0.9707	0.9700	0.9700
SVM	0.9583	0.9689	0.9583	0.9581

According to the model's accuracy of 1.0 for both the validation and test sets, the random forest model is very effective in predicting cancer based on the provided features. It is clear from the confusion matrices for the validation and test sets that there are no false positives or false negatives in the predictions, demonstrating that the model is capable of correctly predicting the outcomes for all three classes. Each class's accuracy, recall, and f1-score are all 1.0, demonstrating that the model can successfully identify every instance of each class. The successful random forest model implies that the chosen features are highly useful in predicting cancer and that the model is able to capture the correlations between the features and the target variable. Based on these findings, it is possible to construct a more reliable diagnostic tool for cancer detection using the random forest model, which may be useful in detecting cancer using these specific traits.

Future work:

In this paper, the Random Forest algorithm was trained, and its hyperparameters were tuned to predict cancer efficiently and accurately. Future work includes collecting more features related to cancer and gather a larger dataset. Firstly, regarding the inclusion of more features related to cancer, it's important to note that cancer is a complex disease with a multitude of factors that can contribute to its development. The current model may not have included all relevant features that could influence an individual's cancer risk. Thus, expanding the feature set could potentially improve the model's performance.

Furthermore, as the dataset used in the current model was relatively small, collecting a larger dataset could also improve the model's accuracy.

Conclusion:

In conclusion, the developed machine learning models for predicting cancer risk based on certain features, and found that Random Forest, Logistic Regression, and KNN performed well. This study provides insight into the potential of machine learning models in predicting cancer risk, which can be useful for early detection and treatment.

References

[1] https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

[2] Varsha Nemade, Vishal Fegade,Machine Learning Techniques for Breast Cancer Prediction,Procedia Computer Science,Volume 218,2023,Pages 1314-1320,ISSN 18770509,https://doi.org/10.1016/j.procs.2023.01.110.(https://www.sciencedirect.com/science/article/pii/S1877050923001102)

[3] R. Uppara, S. Yadav and D. M. Kavitha, "Voting Classifier on Ensemble Algorithms for Breast Cancer Prediction,“ 2023

[4] Shafique, R.; Rustam, F.; Choi, G.S.; Díez, I.d.l.T.; Mahmood, A.; Lipari, V.; Velasco, C.L.R.; Ashraf, I. Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning. Cancers 2023, 15, 681. https://doi.org/10.3390/cancers15030681

[5] Article Source: **Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit**
Huang T, Le D, Yuan L, Xu S, Peng X (2023) Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. PLOS ONE 18(1): e0280606. https://doi.org/10.1371/journal.pone.0280606

[6] Kumar, Ajay and Sushil, Rama and Tiwari, Arvind Kumar, Machine Learning Based Approaches for Cancer Prediction: A Survey (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Available at SSRN: https://ssrn.com/abstract=3350294 or http://dx.doi.org/10.2139/ssrn.3350294

[7] Polygenic risk scores and breast cancer risk prediction Eleanor Roberts a , Sacha Howell a,d,e , D Gareth Evans b,c,d,e,*

[8] AIP Conference Proceedings 2168, 020050 (2019); https://doi.org/10.1063/1.5132477 Published Online: 04 November 2019

[9] Manas Minnoor, Veeky Baths,Diagnosis of Breast Cancer Using Random Forests,Procedia Computer Science,Volume 218,2023,ISSN 1877 0509,(https://www.sciencedirect.com/science/article/pii/S187705092300025X)

[10] Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis Mohammed Amine Naji a, *, Sanaa El Filalib Kawtar Aarikac , EL Habib Benlahmard , Rachida Ait Abdelouhahide , Olivier Debauche

[11] Varsha Nemade, Vishal Fegade,Machine Learning Techniques for Breast Cancer Prediction,Procedia Computer Science,Volume 218,2023,Pages 1314-1320,ISSN 18770509,https://doi.org/10.1016/j.procs.2023.01.110.(https://www.sciencedirect.com/science/article/pii/S1877050923001102)

[12] Kumar, Ajay and Sushil, Rama and Tiwari, Arvind Kumar, Machine Learning Based Approaches for Cancer Prediction: A Survey (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Available at SSRN: https://ssrn.com/abstract=3350294 or http://dx.doi.org/10.2139/ssrn.3350294

[13] Predicting metastasis in gastric cancer patients: machine learning-based approaches AtefehTalebi1,2, CarlosA. Celis-Morales2,3, Nasrin Borumandnia 4*, SomayehAbbasi5 , MohamadAmin Pourhoseingholi6 , AbolfazlAkbari7 & JavadYousef

[14] R. Uppara, S. Yadav and D. M. Kavitha, "Voting Classifier on Ensemble Algorithms for Breast Cancer Prediction,“ 2023

[15] Shih, Yang-Hsin & Qin, Gong & Tang, Peifu & Shen, & Liu, Tai-Yi & Gao, Shuai. (2019). Delineation of Urban Growth Boundaries Using a Patch-Based Cellular Automata Model under Multiple Spatial and Socio-Economic Scenarios. Sustainability. 11. 6159. 10.3390/su11216159.