Assignment 2: Ensemble Machine Learning Model for Cancer Diagnosis Prediction

DSC 415: Data Analysis and Visualization

Nupur Upadhyay
IMS22191
Indian Institute of Science Education and Research, Thiruvananthapuram
Kerala, India
nupur22@iisertvm.ac.in

***Abstract*—This study presents an exploratory data analysis (EDA) and predictive modelling approach for cancer diagnosis using a publicly available Kaggle dataset. EDA was conducted to examine the distribution of features, identify correlations, detect missing values, and uncover patterns relevant to cancer diagnosis. Insights from EDA guided data preprocessing steps. Multiple machine learning models were implemented, ranging from baseline classifiers such as Logistic Regression and Decision Trees to advanced techniques like Random Forest, Gradient Boosting, and other built-in ensemble methods, along with novel meta-ensemble and voting models. Model performance was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and ROC. Comparative analysis revealed that ensemble-based models achieved superior predictive performance, highlighting their robustness in handling complex biomedical data. The results demonstrate the potential of data-driven approaches in supporting cancer prediction and provide a foundation for future work in applying machine learning for medical diagnostics.**

***Index Terms*—datasets, EDA, vizualization, Machine Learning, cancer prediction, diagnosis**

I. Introduction

Cancer is a major global health concern, and early and accurate prediction is critical for improving patient survival and treatment outcomes. Traditional diagnostic methods, while effective, can be time-consuming and resource-intensive, motivating the application of computational approaches. Machine learning (ML) has emerged as a powerful tool for uncovering complex patterns in biomedical data, enabling the development of predictive models that can assist in diagnosis and decision-making.

In this study, exploratory data analysis (EDA) was first performed to investigate feature distributions, correlations, and potential class imbalances within the cancer dataset. Insights from EDA informed pre-processing steps such as normalization and feature selection, which improved the quality of inputs for model training. A variety of ML algorithms were then implemented, ranging from baseline models like Logistic Regression and Decision Trees to advanced ensemble techniques including Random Forest and Gradient Boosting. Then, Voting and Blending approaches were used to create new Ensemble models. Model performance was assessed using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, allowing for a comprehensive comparison. Results demonstrated that ensemble methods consistently outperformed individual models, highlighting their effectiveness in capturing complex relationships in biomedical data and their potential utility in supporting cancer prediction.

II. About the dataset

The dataset used for cancer prediction is Cancer Prediction Dataset by Rabie El Kharoua on Kaggle. This dataset consists of seven features: Age: Integer values representing the patient's age, ranging from 20 to 80. Gender: Binary values representing gender, where 0 indicates Male and 1 indicates Female. BMI: Continuous values representing Body Mass Index, ranging from 15 to 40. Smoking: Binary values indicating smoking status, where 0 means No and 1 means Yes. Genetic Risk: Categorical values representing genetic risk levels for cancer, with 0 indicating Low, 1 indicating Medium, and 2 indicating High. Physical Activity: Continuous values representing the number of hours per week spent on physical activities, ranging from 0 to 10. Alcohol Intake: Continuous values representing the number of alcohol units consumed per week, ranging from 0 to 5. Cancer History: Binary values indicating whether the patient has a personal history of cancer, where 0 means No and 1 means Yes. Diagnosis: Binary values indicating the cancer diagnosis status, where 0 indicates No Cancer and 1 indicates Cancer. This dataset has 1500 samples.

III. Methodology

The cancer prediction model methodology outlines the key components and processes in developing and deploying a hybrid machine learning model to predict cancer diagnosis. Box 1 shows the systemic overview of the prediction model using machine learning.

Box 1: Overview of Cancer Prediction Workflow

Cancer Prediction Dataset
↓
Data Pre-processing
↓
Exploratory Data Analysis
↓
Machine Learning Models
↓
Model Training
↓
Tuning and Optimization
↓
Model Evaluation

A. Data Collection

The dataset used for cancer prediction is Cancer Prediction Dataset by Rabie El Kharoua on Kaggle. The dataset was imported to Google Colab using Kagglehub, and further analyses were performed. The target variable (to be predicted) is the diagnosis feature, which specifies whether a person is diagnosed with cancer or not. The rest of the features (Age, Physical activity, etc.) will be the features used for predicting the diagnosis. Each row refers to a patient who was tested for cancer.

B. Data Pre-processing

Before applying machine learning models, the dataset was pre-processed to ensure accuracy and consistency. This includes handling missing values, encoding categorical features, scaling numerical variables, and removing any duplicates or inconsistencies. This dataset had no missing values, and the categorical features had been pre-encoded. Scaling of the dataset was done using the StandardScaler pre-processing tool from the Scikit-learn library. Then, both the scaled and unscaled datasets were split into training (80 per cent) and testing (20 per cent) sets to enable effective model evaluation.

C. Exploratory Data Analysis

Next, Exploratory Data Analysis (EDA) was performed in Python using visualization libraries such as Seaborn and Matplotlib. Various plots, including histograms, box plots, and heatmaps, were generated to examine the distribution of features, detect outliers, and understand correlations between variables. These visualizations provided insights into data patterns and relationships, which guided further pre-processing and model selection.

D. Machine Leaning Models

Machine learning models were implemented in Python using scikit-learn. After preparing the dataset, several algorithms were applied to train predictive models, including both baseline and advanced approaches. In addition, ensembling techniques such as Blending and Voting (Weighted averaging) were employed to combine the strengths of multiple models, aiming to achieve better predictive performance than individual models.

E. Tuning and Optimization

The models were evaluated on both scaled and unscaled datasets to examine the effect of feature scaling. It was observed that scaling led to lower performance in this case, so the unscaled version of the dataset was retained for further analysis. For ensemble models, tuning involved adjusting base learner parameters as well as meta-model configurations to maximize overall accuracy and reduce overfitting.

F. Model Evaluation

Model performance was evaluated on the testing dataset using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Confusion matrices and performance curves were also analyzed to assess how well the models generalized to unseen data. For ensemble methods, performance improvements were compared against individual base models to highlight the effectiveness of weighted averaging and Blending strategies.

IV. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset before applying machine learning models. This step involved examining the distribution of features, detecting missing values or outliers, and exploring relationships between variables through statistical summaries and visualizations. EDA provided critical insights that guided data pre-processing, feature selection, and overall modelling strategy.

A. Outlier Detection

Outlier detection was performed using box plots created with Seaborn. Box plots provided a clear visualization of the spread of each feature, highlighting the median, quartiles, and extreme values (Figure 1). Data points lying beyond the whiskers are considered as potential outliers; however, there were no outliers in this dataset.

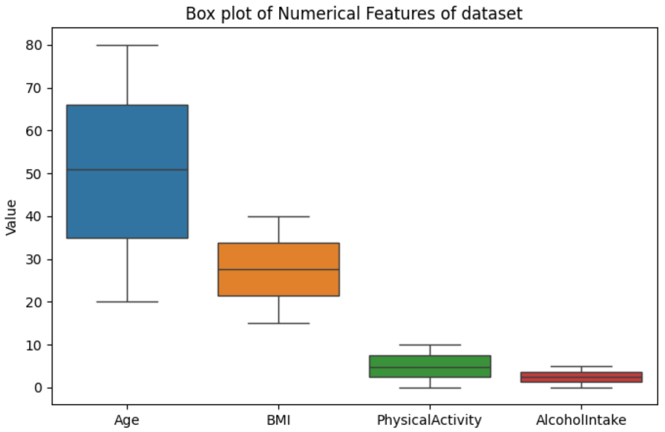
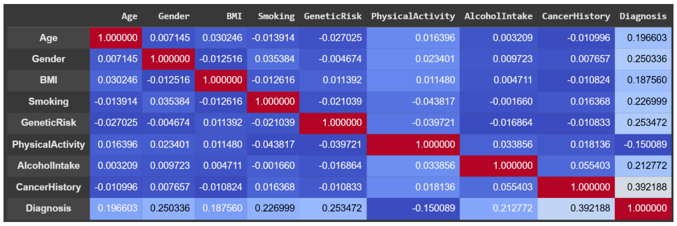


Fig. 1. Outlier Detection

B. Correlation matrix

Then, to know the correlation between different variables of the dataset better, a correlation matrix was plotted and visualized using the ”coolwarm” palette (Figure 2). Here, in the image, the correlation coefficients for each pair of variables are shown. Correlation coefficients with high positive correlation are in red and the ones with more negative in darker blue. All the variables have a weak correlation to the target variable, and thus, it will be difficult to predict the diagnosis using only one variable. Thus, machine learning models become incredibly useful here.



	Age	Gender	BMI	Smoking	GeneticRisk	PhysicalActivity	AlcoholIntake	CancerHistory	Diagnosis
Age	1.000000	0.007145	0.030246	-0.013914	-0.027025	0.016396	0.003209	-0.010996	0.196603
Gender	0.007145	1.000000	-0.012516	0.035384	-0.004674	0.023401	0.009723	0.007657	0.250336
BMI	0.030246	-0.012516	1.000000	-0.012616	0.011392	0.011480	0.004711	-0.010824	0.187560
Smoking	-0.013914	0.035384	-0.012616	1.000000	-0.021039	-0.043817	-0.001660	0.016368	0.226999
GeneticRisk	-0.027025	-0.004674	0.011392	-0.021039	1.000000	-0.039721	-0.016864	-0.010833	0.253472
PhysicalActivity	0.016396	0.023401	0.011480	-0.043817	-0.039721	1.000000	0.033856	0.018136	-0.150089
AlcoholIntake	0.003209	0.009723	0.004711	-0.001660	-0.016864	0.033856	1.000000	0.055403	0.212772
CancerHistory	-0.010996	0.007657	-0.010824	0.016368	-0.010833	0.018136	0.055403	1.000000	0.392188
Diagnosis	0.196603	0.250336	0.187560	0.226999	0.253472	-0.150089	0.212772	0.392188	1.000000

Fig. 2. Correlation Matrix

C. Sample Distribution based on Sex

To find out how the samples were distributed across this dataset, a bar plot was plotted for the count of males and females in this sample (Figure 3). It can be seen that the count of samples is almost equal for both sexes. This verifies that the dataset is a good representative for both sexes.

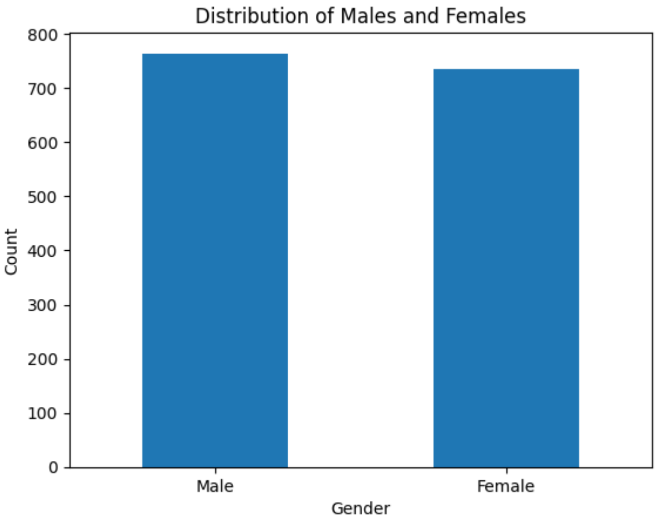


Fig. 3. Sample distribution between males and females

D. Age wise distribution

Then, to explore the age range of the patients and the number of patients in each age group, a histogram was plotted (Figure 4). The age range is between 20 and 80, and there is not a big difference between the sample counts for each age range. Thus, the data is not biased towards any age group.

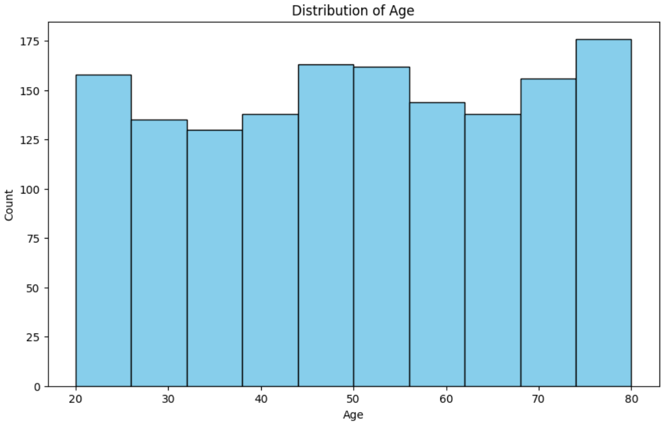


Fig. 4. Age distribution

E. Physical Activity Distribution

Similarly, to know about the distribution of the physical activity (Number of hours spent on physical activity per week), a histogram was plotted. The dataset contains a distribution of samples ranging from not active (0 hours) to active (10 hours). The distribution is not skewed towards any group.

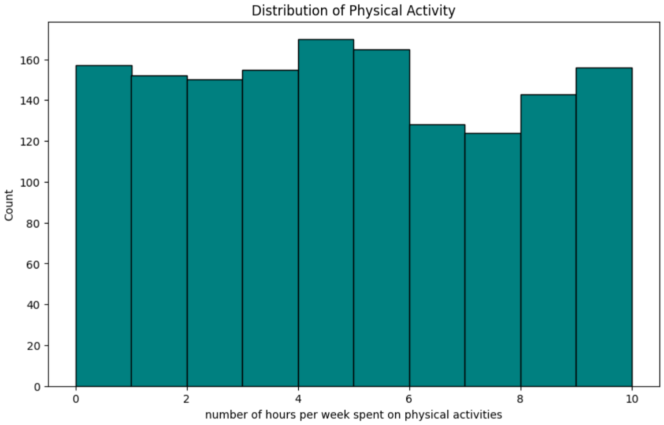


Fig. 5. Physical Activity Distribution

F. Cancer diagnosis

Next, to check the distribution of positive and negative diagnoses for cancer, a pie chart was plotted (Figure 6). The percentage of positive diagnoses is less, and thus, only 37.1 per cent of the samples were diagnosed as positive.

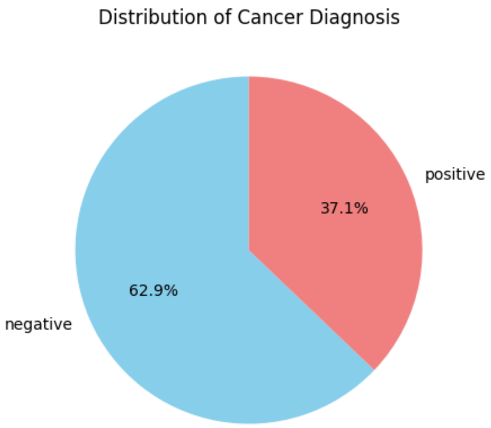


Fig. 6. Cancer Diagnosis

G. *Cancer History*

Then, to check how many people had been diagnosed with cancer before, a pie chart was plotted (Figure 7). Only a small fraction of people have a history of Cancer.

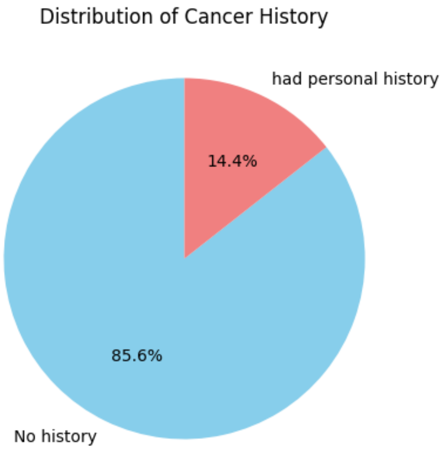


Fig. 7. Cancer History distribution

H. *Cancer Diagnosis and History*

Next, to explore how a person's personal history with cancer affects the diagnosis outcome, a bar plot was plotted with only people with a personal history of cancer under consideration (Figure 8). From the graph, it can be concluded that personal history doesn't affect the diagnosis a lot, as only a small fraction of people with personal history get diagnosed with cancer.

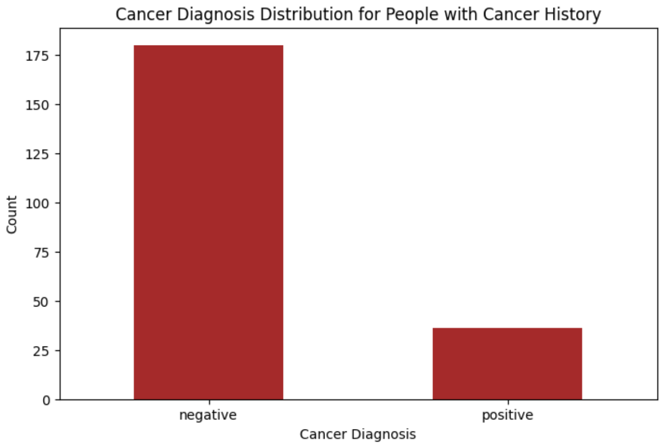


Fig. 8. Cancer diagnosis for people with personal Cancer History

V. Machine Learning Models

To carry out the prediction for cancer diagnosis using built-in ensemble models, voting ensemble and meta-ensemble machine learning models, the scikit-learn library was used. The train test split function from scikit-learn was used with a test size of 0.2 to create the train and test datasets. Data scaling was done using the StandardScaler tool. However, the performance of the ML models was reduced by a significant margin when trained and tested on the scaled data. Thus, only unscaled data was used for model training, testing and evaluation. The standard machine learning models trained were: Logistic Regression, Decision Tree, Random Forest, k-NN (k Nearest Neighbours), SVM (Support Vector Machine), Gradient Boosting and MLP (Multi-Layer Perceptron). In addition, two ensemble models were trained: A Voting Ensemble (Weighted Average) and a Blending Ensemble of the two high-performing models- Random Forest and Gradient Boosting Algorithm.

A. *Evaluation Results*

*1) **Built-in Models**:* The trained models were then tested on the test (20 per cent) dataset. The results for the evaluation metrics, the Confusion Matrix and the ROC curve for each Baseline model are as follows.

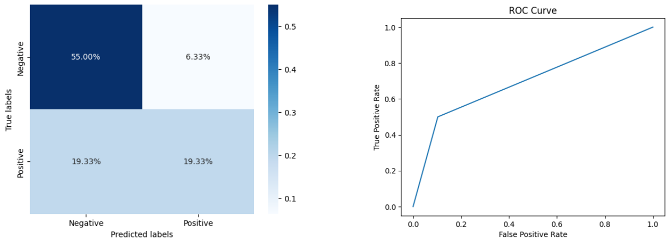


Fig. 9. Logistic Regression

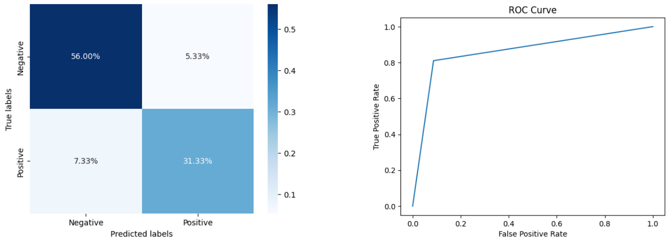


Fig. 10. Decision Tree

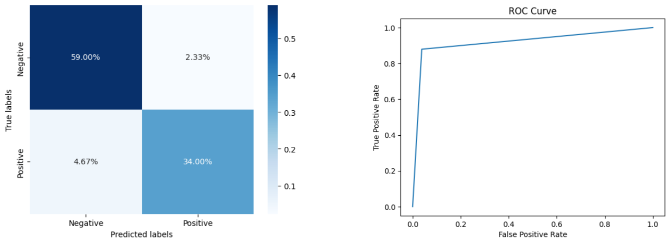


Fig. 11. Random Forest

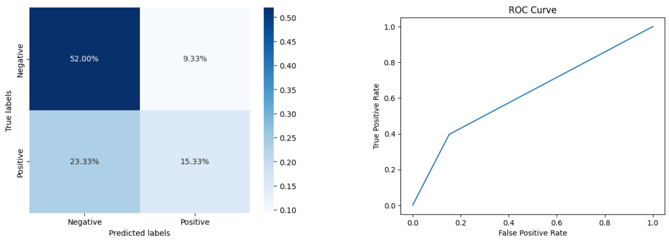


Fig. 12. k-NN

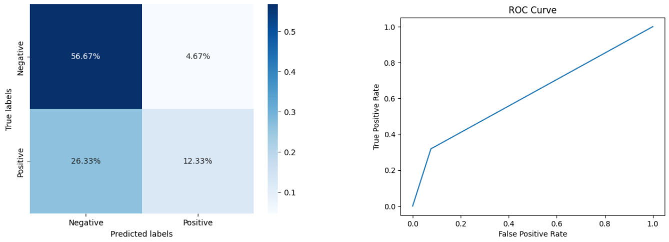


Fig. 13. SVM

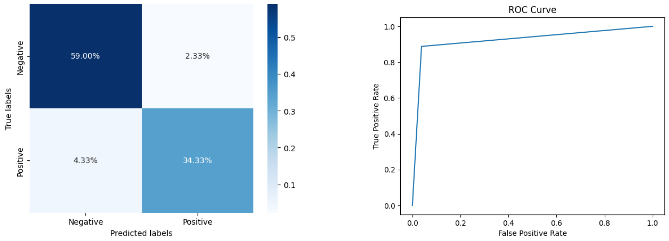


Fig. 14. Gradient Boosting

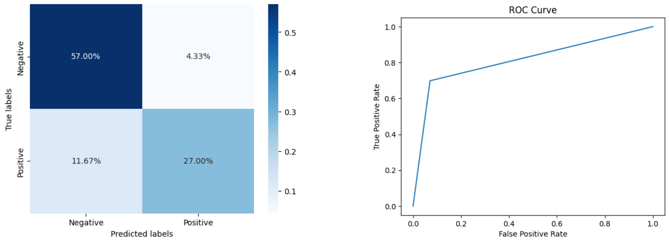


Fig. 15. MLP

*2) **proposed Ensemble models:*** Blending is an ensemble learning technique in which predictions from multiple base models are combined using a meta-learner trained on a separate hold-out set. Unlike stacking, which relies on cross-validation to generate out-of-fold predictions, blending employs a fixed validation (blending) set to reduce information leakage. In this work, the training data was partitioned into two subsets: a base-training set used to fit Random Forest and Gradient Boosting classifiers, and a hold-out blending set on which these models generated predicted probabilities. These predictions were then stacked into a new feature space and used to train a Logistic Regression meta-learner, enabling the ensemble to leverage complementary decision boundaries of the base models for improved predictive performance.

Weighted averaging is an ensemble method that combines the predicted probabilities of multiple base models by assigning predefined weights to each model. In this study, predictions from the Random Forest and Gradient Boosting classifiers were averaged with equal weights, and the resulting probabilities were thresholded to obtain the final class predictions. This approach provides a simple yet effective way to integrate the strengths of individual models without training a separate meta-learner.

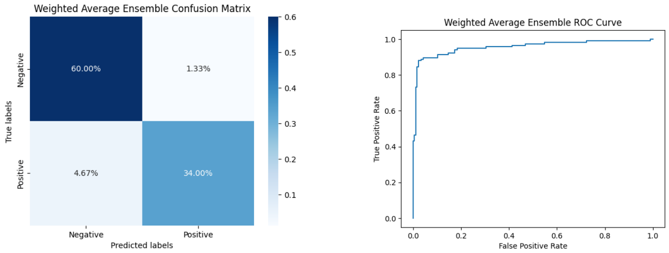


Fig. 16. Weighted Average Ensemble

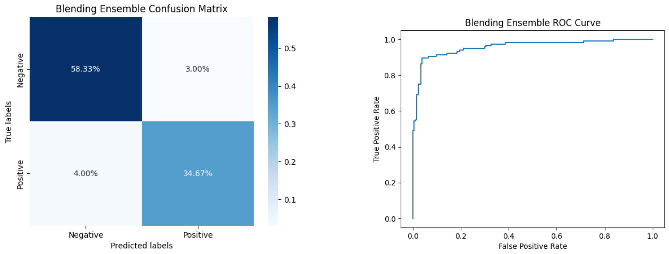


Fig. 17. Blending Ensemble

Thus, both the voting and blending ensemble models performed well on the test data, with the Weighted Average model slightly outperforming the Blending model. In conclusion, built-in Ensemble models like Gradient Boosting, Random Forest and the Blending Ensemble model performed well with an accuracy and precision score of 93 per cent. However, the Weighted Average Ensemble model outperformed all these models with an Accuracy and Precision score of 94 per cent.

The overall accuracy, precision, recall and F1 score of the models are given in Figures 18, 19, 20 and 21, respectively. The models with the metric of more than 90 per cent are highlighted in green.

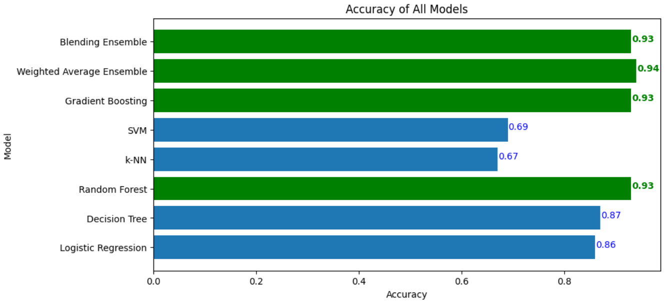


Fig. 18. Accuracies of the ML models

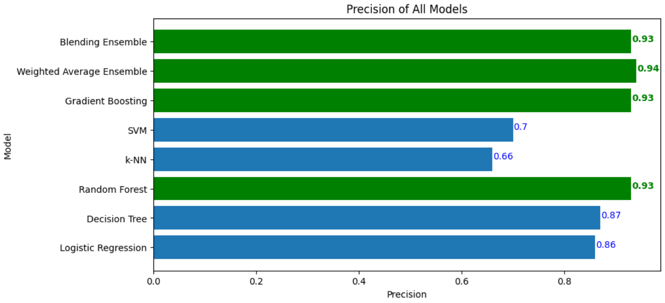


Fig. 19. Precision of the ML models

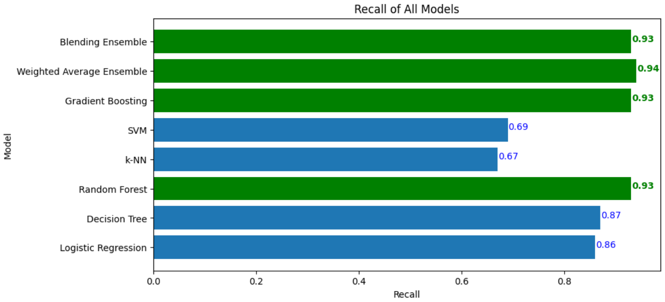


Fig. 20. Recall of the ML models

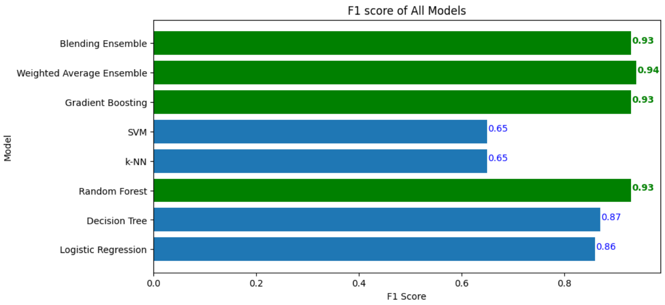


Fig. 21. F1 score of the ML models

The overall metrics for all the models are summarized in Table 1.

TABLE I
EVALUATION OF THE ML MODELS

Machine Learning	Evaluation Metrics			
Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.86	0.86	0.86	0.86
Decision Tree	0.87	0.87	0.87	0.87
Random Forest	0.93	0.93	0.93	0.93
k-NN	0.67	0.66	0.67	0.65
SVM	0.69	0.7	0.69	0.65
Gradient Boosting	0.93	0.93	0.93	0.93
MLP	0.84	0.84	0.84	0.84
Weighted Average Ensemble	0.94	0.94	0.94	0.94
Blending Ensemble	0.93	0.93	0.93	0.93

The ROC curves of all the models, along with AUC (Area under the curve) are summarized in Figure 22.

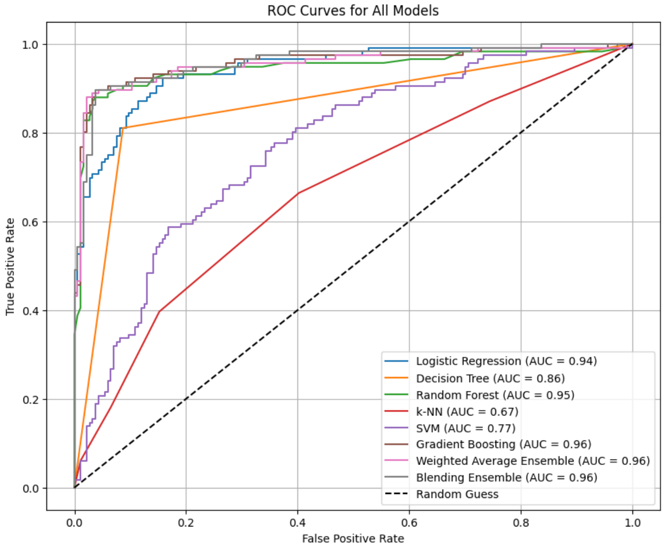


Fig. 22. ROC curves of ML models

Thus, while some built-in Ensemble models like Random Forest and Gradient Boosting are good at predicting the target variable (cancer diagnosis), the Blending Ensemble outperforms them by a small margin. However, in clinical settings where the outcome of the diagnosis prediction could be very critical to the patient's health, even a marginal improvement in prediction could be really meaningful. Thus, the Blending Ensemble of Random Forest and Gradient Boosting can be used as a prediction tool for cancer diagnosis.

In datasets like these, where there is no strong correlation between the features and the target variable, it makes it difficult to use the feature variables as a marker for prediction. In cases like these, Machine Learning models can achieve a high accuracy in prediction and thus could be a very useful tool in addition to the usual medical diagnostic tools.