# Predicting and Classifying Perceptions of Learning Needs Importance among Patients with Cancer: A Machine Learning Approach


## Abstract

**Background:** Artificial intelligence (AI) and machine learning (ML) are revolutionizing healthcare by enhancing the prediction of learning needs and enabling tailored educational interventions for patients and staff.This study explores the application of AI and ML models to predict learning needs from the patient's perspective.

**Methods:** Three ML models—Linear Regression, Random Forest, and Gradient Boosting—were trained on health literacy, demographic, and treatment data from 218 cancer patients at Sultan Qaboos Comprehensive Cancer Center. Evaluation metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), $R^2$ Score, and Area Under the Curve (AUC). Classification models (Random Forest, Gradient Boosting, Decision Tree, and Extra Trees) were assessed for Accuracy, Precision, Recall, F1-score, and AUC in categorizing learning needs.

**Results:** Gradient Boosting had the best predictive performance (MAE: 0.0534, RMSE: 0.0788, $R^2$: 0.9844, AUC: 0.96), followed by Random Forest (AUC: 0.93). Linear Regression was less effective (AUC: 0.85). Key predictors included literacy level in chemotherapy, hormonal therapy, and treatment experiences, while demographic factors had minimal impact. For classification, Gradient Boosting and Decision Tree models achieved the highest accuracy (96.51%) and AUC (0.96). Random Forest showed 94.19% accuracy, while Extra Trees had 90.70%, indicating variability in model performance.

**Conclusion:** AI and ML, particularly Gradient Boosting, demonstrate strong potential in predicting and categorizing learning needs. These models enable targeted educational strategies, addressing knowledge gaps and aligning interventions with treatment experiences to improve healthcare quality.



**Keywords:** Artificial intelligence, machine learning, learning needs, healthcare education, predictive modeling, health literacy, personalized interventions


## Introduction

Machine learning (ML) and artificial intelligence (AI) are at the forefront of technological advancements in healthcare, playing pivotal roles in predicting learning needs and evaluating their perceived importance [1,2]. These technologies leverage the power of data to create predictive models that highlight gaps in knowledge and skills, thereby empowering educators and healthcare managers to develop more effective, patient-centered learning programs. By analyzing large and complex datasets, ML algorithms can identify trends and insights that would otherwise remain hidden, helping healthcare providers and patients address specific deficiencies in knowledge and skill development [1,2].

One of the key advantages of ML models is their ability to monitor training outcomes in real-time. Unlike traditional methods of education evaluation, which may rely on periodic assessments, ML enables continuous adaptation of learning programs. For instance, when healthcare staff or patients demonstrate improvement in certain areas but show deficits in others, ML can dynamically adjust the educational content to target these emerging gaps [3]. Innovations such as federated learning have further enhanced this capability by allowing collaboration across multiple organizations while maintaining stringent data privacy standards. This ensures that learning needs are identified and prioritized on a larger scale without compromising sensitive information [3, 4].

The application of ML and AI extends beyond simple identification of knowledge gaps. These tools are now being used to assess and integrate the **perceived importance** of these learning needs, a factor that significantly influences engagement, satisfaction, and the overall effectiveness of educational interventions. For example, understanding how patients and healthcare providers prioritize different aspects of education allows for a more targeted approach to designing programs that resonate with their expectations and goals [4,5]. This ensures that the educational resources are not only comprehensive but also aligned with the specific needs and preferences of the audience [5].

AI has been particularly transformative in the context of chronic disease management, including cancer care, where it helps predict learning needs and their perceived importance. For example, in the management of diabetes and hypertension, ML algorithms have been successfully used to identify

gaps in clinical knowledge and skills among healthcare providers, guiding the development of focused training programs [6]. Similarly, in oncology, AI tools predict learning needs related to disease progression, treatment side effects, and psychological support, while simultaneously assessing how patients prioritize these areas of education. This dual focus ensures that educational interventions are tailored not only to address objective knowledge gaps but also to align with patients' perspectives and priorities [6,7].

Cancer patients often face complex and multifaceted challenges that necessitate tailored education to navigate their diagnosis, treatment options, and self-management strategies. The perception of these learning needs—how important patients consider specific topics—plays a critical role in determining their engagement with educational programs. For instance, patients who perceive their learning needs as unmet may experience anxiety, frustration, or disengagement, negatively impacting their adherence to treatment plans and overall health outcomes. Conversely, addressing learning needs that patients deem important fosters a sense of empowerment, improves their quality of life, and enhances their satisfaction with care [8,9].

AI and ML models are invaluable in assessing the perceived importance of learning needs. By analyzing diverse sources such as patient feedback, demographic data, and behavioral patterns, these tools can identify trends in how patients prioritize educational topics. For instance, individuals with low literacy levels may value simplified resources that explain treatment protocols in layman's terms, while those with higher literacy levels may seek detailed information about advanced therapies or clinical trials [10,11]. This segmentation enables healthcare providers to create personalized educational materials that not only fill knowledge gaps but also resonate with the preferences and expectations of individual patients.

The integration of AI-driven insights allows healthcare providers to address not only the cognitive aspects of learning but also the emotional and psychological dimensions. For example, patients undergoing chemotherapy may prioritize learning about managing side effects, while others may find psychological support and coping mechanisms more critical. AI tools can assess these preferences and provide tailored resources that enhance patients' trust in their care teams and facilitate active participation in their treatment plans. This approach ensures that education is holistic, addressing the physical, emotional, and informational needs of patients [12,13].

AI technologies are increasingly being applied to identify and address learning needs in cancer care. For instance, ML algorithms analyze patient-reported outcomes and satisfaction surveys to detect gaps in education related to treatment protocols, genetic testing, and symptom management. These tools predict the importance patients assign to specific learning topics, enabling the development of targeted educational resources that cater to both objective needs and subjective priorities [14,15].

In addition, AI's role in cancer education extends to analyzing intervention outcomes in mental healthcare. Similar methodologies can be applied in oncology to assess patients' priorities and provide tailored resources that enhance their understanding of disease management and self-care [16,17]. Predictive models generated by AI not only personalize the learning experience but also ensure that it evolves with the patient's changing needs and preferences, creating a dynamic and responsive educational framework [14,15].

Innovative approaches, such as federated learning, enhance these applications by facilitating collaboration among healthcare institutions while preserving patient privacy. This enables the pooling of diverse data sources, leading to a more comprehensive understanding of collective learning needs and educational priorities across different populations [3,18].

The perception of learning needs and their importance directly influences patient satisfaction, engagement, and health outcomes [19]. Patients who feel that their educational needs are prioritized are more likely to engage actively in their care, adhere to treatment protocols, and achieve better health outcomes. By leveraging AI to understand these perceptions, healthcare providers can design interventions that address not only knowledge gaps but also the emotional and psychological priorities of patients. This alignment fosters trust, empowers patients, and ultimately improves the overall quality of care [19,20].

This study explores the application of AI and ML models in predicting and addressing the perceived importance of learning needs from the patient's perspective, with a focus on oncology. By integrating advanced predictive technologies with patient-centered education strategies, this research demonstrates the potential of AI to transform cancer care. Through a combination of objective data analysis and subjective prioritization, AI ensures that educational interventions are relevant, effective, and aligned with the unique needs of patients.

Methods

This study employed a machine learning approach to predict and classify perceptions of the importance of learning needs. The methods consisted of data collection, preprocessing, exploratory data analysis, feature engineering, model training, and evaluation. The focus was on utilizing machine learning models to predict and classify learning needs based on various features, including demographic data, health literacy scores, and experience metrics.

Setting and Design

The study was conducted at the Sultan Qaboos Comprehensive Cancer Center, University Medical City, located in Muscat, Oman. A cross-sectional design was conducted to assess health literacy, learning needs among cancer patients.

Sampling

The population for this study consisted of cancer patients who were receiving treatment at the Sultan Qaboos Comprehensive Cancer Center (SQCCCRC) in Muscat, Oman. Participants of the study needed to meet the following inclusion criteria:

- Diagnosed with cancer.
- Receiving treatment at the SQCCCRC.
- Willingness to participate and able to complete the survey online.

A convenience sampling approach was employed to select participants for the study to recruit a representative sample size of 218 patients. This sample size had been selected based on a 95% confidence level and a 5% margin of error.

Instrument:

To comprehensively capture the study's objectives, a self-reported questionnaire was employed, comprising the following sections:

- **Demographic Information:** Included age, gender, region, job status, marital status, diagnosis, diagnosis date, and current treatment.
- **Health Literacy Assessment:** The Health Literacy Instrument for Adults (HELIA), which consists of 33 items assessing reading, access to information, understanding, appraisal, and decision-making, was used. It ranks health literacy from 0 (inadequate) to 100 (excellent).(17)
- **Needs Assessment Questionnaire:** This survey covered psychosocial, informational, physical, patient care, and communication domains. Participants rated each domain's importance on a 5-point Likert scale. Items were derived from Chua et al.[18]

Data Collection

After obtaining approval from the Institutional Review Board (IRB) at the Sultan Qaboos Comprehensive Cancer Center (SQCCRC), potential participants were recruited face-to-face by the research team using an information statement. If the patient agreed to participate, an invitation letter with the information statement was sent through WhatsApp. Participants who agreed to participate completed and submitted the self-administered questionnaire.

Data Preprocessing

The collected data underwent extensive preprocessing to prepare it for model training. The preprocessing steps included data cleaning, normalization, feature encoding, and data splitting:

- Data Cleaning: Mean imputation was used for numerical characteristics and mode imputation was used for categorical data in order to handle missing values. Interquartile range (IQR) analysis was used to identify outliers, which were then either eliminated or adjusted to guarantee data quality.
- Normalization: In order to make sure that all numerical features were scaled to have comparable ranges—a crucial step for distance-based models—normalization was applied.
- Feature Encoding: To prepare them for machine learning models, categorical variables—like gender and educational attainment—were encoded using one-hot encoding. In order to maintain their inherent order, ordinal features—like health literacy levels—were also converted into numerical values.
- Data Splitting: Following preprocessing, an 80-20 split ratio was utilized to assign the dataset into training and testing sets. This division made sure that 80% of the data was utilized for

training the models and 20% was set aside for testing the models' performance on data that had not yet been seen.

Exploratory Data Analysis (EDA)

Prior to model training, an exploratory analysis was utilized out to learn the collected data. Visualizing feature distributions, analyzing variable correlations, and spotting patterns in the dataset were all part of EDA. Understanding the distribution of learning needs across various demographic groups and identifying substantial relationships between health literacy scores and the relevance of learning needs were among the key results.

Feature Engineering

To increase the models' capacity for prediction, feature engineering was done. As part of this process, new features were created from the ones that already existed. For example, aggregated health literacy scores and interaction terms between various features (such as those between demographic characteristics and experience metrics) were calculated. The most significant characteristics that contributed to the target variable were found using feature selection approaches, such as mutual information analysis and recursive feature elimination (RFE), which decreased dimensionality and enhanced model performance.

Prediction Models

Three machine learning models—Gradient Boosting Regressor, Random Forest Regressor, and Linear Regression—were used to make predictions. The models were selected because they each used a different approach to identifying patterns in the data.

- Linear Regression: Because of its ease of interpretability and use, it served as the baseline model. It is a useful model for preliminary analysis since it presumes a linear relationship between the features and the target variable.
- Random Forest Regressor: Because of its capacity to manage non-linearity and feature interactions, Random Forest, an ensemble model, was selected. It is composed of several decision trees, and by averaging them, overfitting is decreased, improving performance.
- Gradient Boosting Regressor: By gradually constructing an ensemble of weak learners and optimizing for decreased error in each iteration, Gradient Boosting was utilized to improve prediction accuracy. This approach is renowned for its capacity to identify intricate connections within the data.

Grid search cross-validation was used for hyperparameter tuning, which optimized each model's parameters, including the maximum decision tree depth and the number of estimators. To reduce the chance of overfitting, model performance was evaluated during tuning using five-fold cross-validation.

The models were evaluated based on multiple metrics to provide a comprehensive assessment of their performance:

• Mean Absolute Error (MAE): Without taking direction into account, MAE was utilized to calculate the average magnitude of the forecast errors. Better model performance was indicated by lower MAE values.

The standard deviation of the prediction errors was measured using the Root Mean Squared Error (RMSE). RMSE is a helpful indicator for identifying notable deviations since it assigns a higher weight to larger errors.

• R2 Score: This metric was used to calculate the percentage of the target variable's variance that the model could account for. A greater percentage of the variance might be explained by the model, according to a higher R2 Score.

To determine which features had the greatest influence on the prediction, feature importance analysis was also carried out for every model. Finding the features in the dataset that had the most effects on the target variable was made easier by the importance scores for features like Total Learning Need Assessment, Total Surgery, and Health Literacy Total Score. Furthermore, to evaluate model performance qualitatively, visualizations of the actual versus anticipated values were made. Scatter plots were used to illustrate how closely the predicted and actual values matched.

Classification Models

Learning needs were categorized using four models: Random Forest, Gradient Boosting, Decision Tree, and Extra Trees, in addition to regression analysis. Using categorical goal variables to provide varying degrees of relevance to learning demands, the models were trained on the same dataset.

• Random Forest: This ensemble model was chosen due to its resilience to overfitting and capacity to manage a large number of features. It improved generality by averaging several decision trees.
• Gradient Boosting: This technique was selected due to its ability to capture intricate feature relationships and handle unbalanced datasets. It constructs trees one after the other, each one attempting to fix the mistakes of the one before it.
• Decision Tree: Because of its interpretability and capacity to manage both numerical and categorical variables, a decision tree was employed. It shed light on the model's decision-making procedure.
• Extra Trees: Using random splits for node splitting, Extra Trees, a Random Forest variant, was used to lower variance and enhance generalization.

Four important measures were used to assess the classification models' performance: accuracy, precision, recall, and F1-score. To provide further light on each model's classification performance and misclassification trends, confusion matrices were created. The confusion matrices showed the distribution of mistakes across classes, and these metrics assisted in evaluating how effectively each model matched sensitivity and specificity.ROC curve study for several classifiers, such as Random Forest, Gradient Boosting, Decision Tree, and Extra Trees, shows how well they predict the "Classification of Importancy." The Area Under the Curve (AUC) score, which shows how effectively the model can differentiate between "Very High Importance" and other classes, is used to evaluate each classifier's performance.

## Results

### Demographics and Variables Information

The table 1 presented demographics and clinic characteristics. In terms of **age**, the sample ranged from 19 to 86 years, while the average was 45.81 years with standard deviation of 15.55 years. In the gender category, 182 patients (56.35%) are female. The largest group consisted of patients with a secondary school education (n = 88, 27.24%). For occupation, most patients were employees (n = 116, 35.91%) and married (n=239, 73.99%) are married. In terms of clinical characteristics, the most patients were diagnosed with rare tumors (n = 107, 33.13%) including sarcoma. Regarding time since diagnosis, the most patients had been diagnosed for over a year (n = 193, 59.75%) and on treatment (n = 197, 60.99%). Lastly, miscellaneous treatments were the most common treatment modalities (n = 128, 39.63%).

Table 1: demographics and clinic characteristics

| **Category** | **Variable** | **Mean (Range)** | **SD** |
|---|---|---|---|
| Age | | 45.81 (19 to 86) | 15.55 |
| | | **Frequency** | **Percentage** |
| Gender | Female | 182 | 56.35% |
| | Male | 141 | 43.65% |
| Education Level | Other | 2 | 0.62% |
| | Primary School | 52 | 16.10% |
| | Secondary School | 88 | 27.24% |
| | Diploma Degree | 60 | 18.58% |
| | Bachelor Degree | 73 | 22.60% |
| | Master Degree | 23 | 7.12% |
| | Doctorate | 4 | 1.24% |
| Occupation | Business Man | 1 | 0.31% |
| | Employee | 116 | 35.91% |
| | House wife | 1 | 0.31% |
| | Housewife | 24 | 7.43% |

| | | | |
|---|---|---|---|
| | Retired | 67 | 20.74% |
| | Student | 18 | 5.57% |
| | Unable to work | 1 | 0.31% |
| | Unemployed | 94 | 29.10% |
| | Widow | 1 | 0.31% |
| Marital Status | Divorced | 9 | 2.79% |
| | Married | 239 | 73.99% |
| | Single | 55 | 17.03% |
| | Widow | 20 | 6.19% |
| Cancer Type | Breast Cancer | 67 | 20.74% |
| | Gastrointestinal Cancer | 63 | 19.50% |
| | Head, Neck and Thoratic Cavity Cancer | 14 | 4.33% |
| | Rare Tumors | 114 | 35.30% |
| | Urinary Tract Cancer | 29 | 8.98% |
| | Women Health Cancer | 25 | 7.74% |
| Time Since Diagnosis | Others | 9 | 2.79% |
| | <3 months | 26 | 8.05% |
| | 3-12 months | 95 | 29.41% |
| | > 1 year | 193 | 59.75% |
| Treatment Status | Newly diagnosed | 18 | 5.57% |
| | Off treatment | 108 | 33.44% |
| | On treatment | 197 | 60.99% |
| Treatment Modalities | Chemotherapy | 77 | 23.84% |
| | Follow-up | 4 | 1.24% |
| | Hormonal | 19 | 5.88% |
| | Immunotherapy | 5 | 1.55% |
| | Miscellaneous | 128 | 39.63% |
| | Radiation | 45 | 13.93% |
| | Surgery | 45 | 13.93% |

Table 2 presented mean scores and standard deviations (SD) for Health Literacy, Learning Need Assessment (Importance. In the Health Literacy variable, the highest mean score was for "Understanding" domains (mean = 4.43, SD = 0.61), meaning that patients generally found this area to be the most developed. The lowest score was for "Appraisal" (mean = 4.00, SD = 0.84). The average total score for health literacy was 4.36 (SD = 0.56), indicating a high level of health literacy among patients. For Learning Need Assessment (Importance), "Chemotherapy/Hormonal Therapy" domain had the highest importance score (mean = 4.65, SD = 0.74). "Clinical Trials" domain had the lowest importance (mean = 4.15, SD = 1.06). The total score for learning need assessment was 4.53 (SD = 0.60), showing a high perception of learning needs importance in most areas.

Table 2: Health literacy, learning need assessment importance, satisfaction with education activities (mean, SD)

| **Category** | **Variables** | **Mean** | **SD** |
|---|---|---|---|
| **Health Literacy** | Reading | 4.29 | 0.78 |

| | | | |
|---|---|---|---|
| | Access | 4.33 | 0.72 |
| | Understanding | 4.43 | 0.61 |
| | Appraisal | 4.00 | 0.84 |
| | Total Score | 4.36 | 0.56 |
| **Total Learning Need Assessment (Importance)** | Diagnosis | 4.38 | 0.72 |
| | Tests and Investigations | 4.54 | 0.82 |
| | Surgery | 4.57 | 0.82 |
| | Radiation Therapy | 4.56 | 0.85 |
| | Chemotherapy/Hormonal Therapy | 4.65 | 0.74 |
| | Clinical Trials | 4.15 | 1.06 |
| | Sexual aspect of Care | 4.34 | 0.81 |
| | Psychosocial aspect of care | 4.54 | 0.71 |
| | Supportive Care | 4.57 | 0.66 |
| | Overall Experience | 4.64 | 0.56 |
| | Total Score | 4.53 | 0.60 |

Prediction Models

The figure 1 presents the evaluation metrics for three regression models: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor, using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and $R^2$ Score as metrics to assess performance.

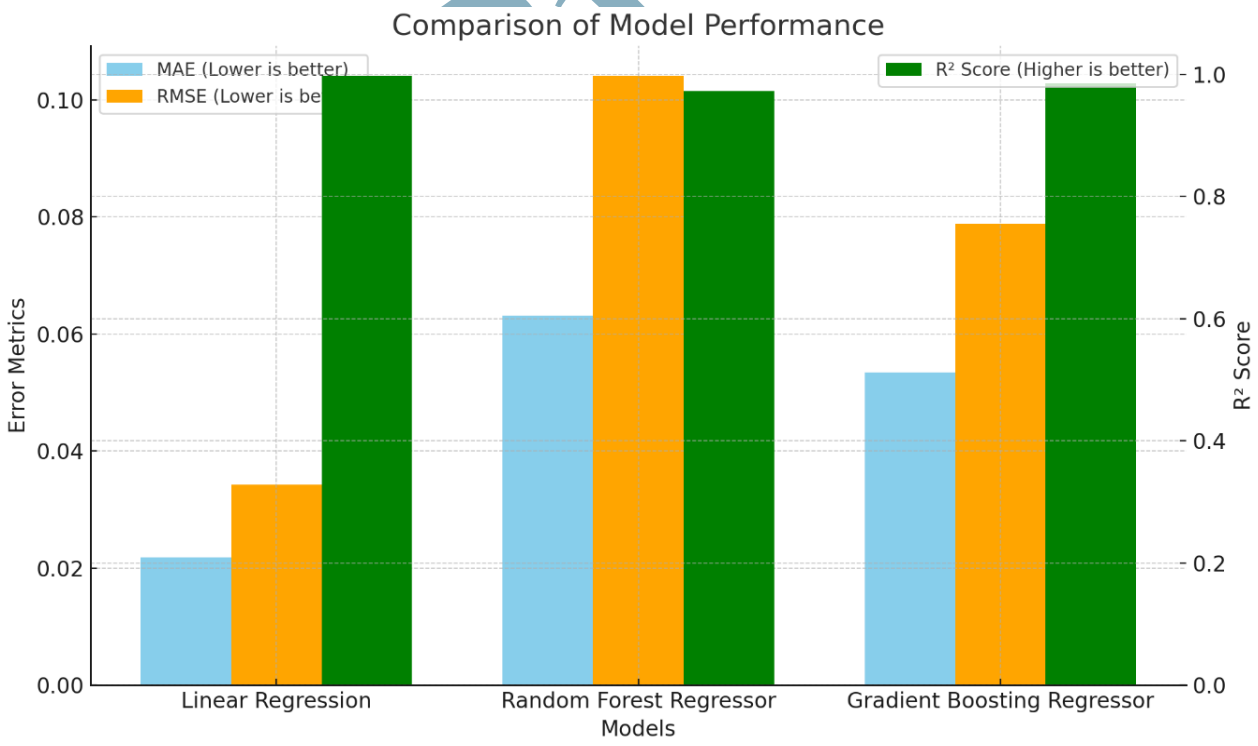


Figure 1: Comparison of model performance

Linear Regression performed the best in terms of MAE and RMSE, with values of 0.0218 and 0.0343, respectively. This indicates that the average and squared deviations of the predicted values from the actual values were the lowest for this model, meaning it had the smallest error among the three

models. Additionally, the $R^2$ Score of 0.997 indicates that the Linear Regression model explained 99.7% of the variance in the target variable, making it the most effective in fitting the data.
Random Forest Regressor had a MAE of 0.0631 and RMSE of 0.1041, which were higher than those of Linear Regression. This implies that the model made slightly larger errors in predicting the target variable. The $R^2$ Score of 0.9727 indicates that it explained 97.27% of the variance, which is still quite good but slightly lower compared to Linear Regression and Gradient Boosting.
Gradient Boosting Regressor achieved MAE and RMSE values of 0.0534 and 0.0788, respectively, which were better than those of Random Forest but not as low as Linear Regression. The $R^2$ Score of 0.9844 suggests that it explained 98.44% of the variance, placing it between Linear Regression and Random Forest in terms of model fit. Overall, Linear Regression demonstrated the lowest error rates and the highest ability to explain the variance in the data, making it the best performer among the three models based on these metrics. Gradient Boosting also performed well, especially in minimizing errors and explaining variance, while Random Forest showed slightly higher error rates but still provided a solid performance.

Feature Importances
The visualization depicts the feature importance scores for three different machine learning models to indicate the level of feature that contribution to the model's predictions, helping us understand which factors play the most significant roles in learning needs importance.

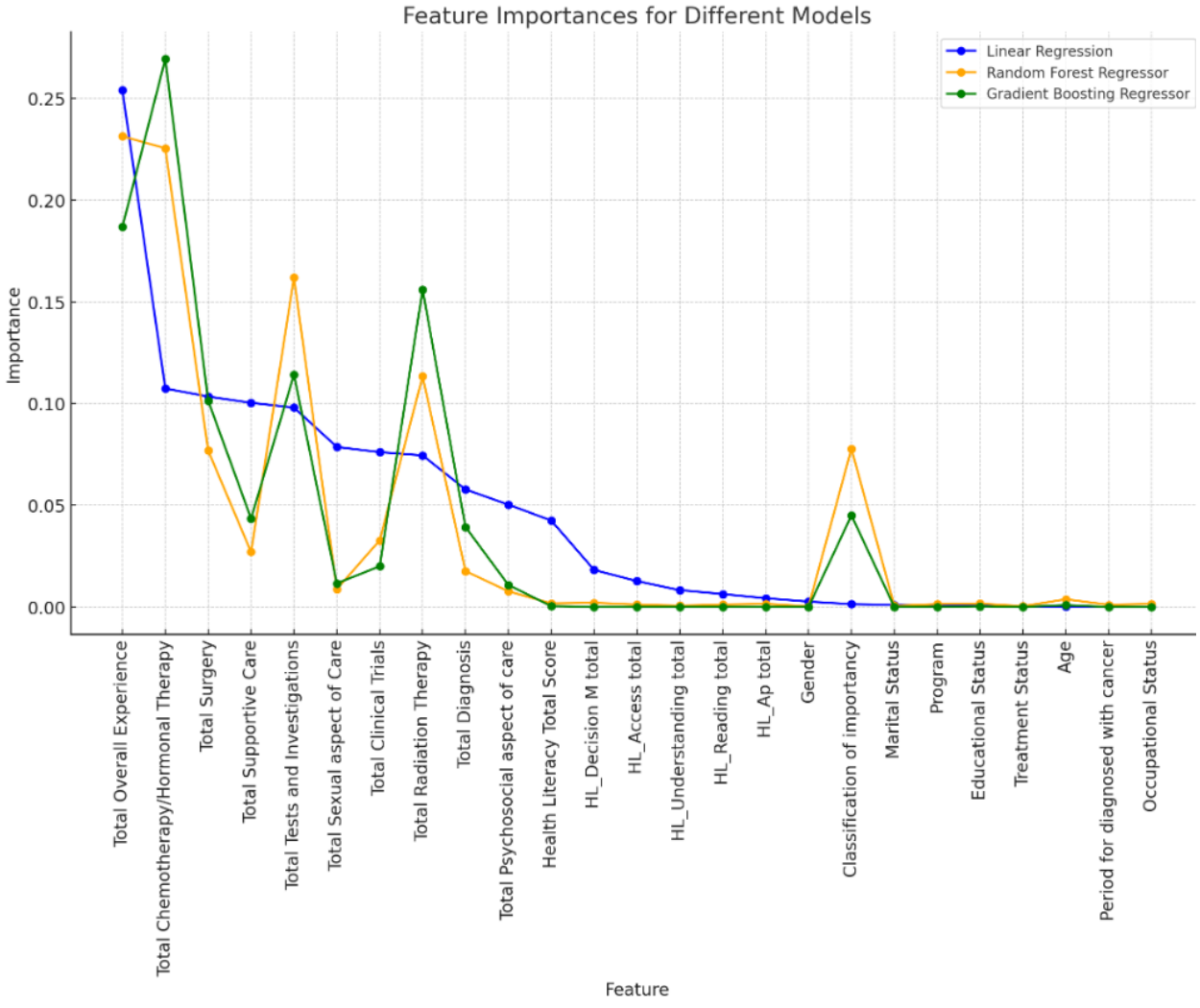


Figure 2: feature importance scores for three different machine learning models: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor

From the chart, Total Chemotherapy/Hormonal Therapy and Total Overall Experience emerged as the most influential features across all models, although the degree of importance varies by model. These features are particularly significant in Gradient Boosting and Random Forest, which assign higher importance to them compared to other features. These findings suggest that the patient's overall experience and specific treatments such as chemotherapy or hormonal therapy are critical in understanding and predicting learning needs.

Total Tests and Investigations and Total Radiation Therapy also show considerable importance, especially in Random Forest and Gradient Boosting. These features contribute substantially to the predictive capabilities of the models, likely reflecting the critical role of diagnostic and therapeutic interventions in shaping patients' perceptions of learning needs.

In contrast, features related to Health Literacy and some demographic characteristics, such as gender and educational status, generally have lower importance scores across all models. This indicates that while these factors might still contribute to the model's understanding, they have less direct impact on learning needs compared to treatment-related experiences and overall patient care.

Overall, the analysis highlights the importance of focusing on patients' treatment experiences and specific medical interventions when predicting and classifying learning needs. Gradient Boosting appears to prioritize fewer but more influential features, while Random Forest distributes the importance more evenly, reflecting each model's distinct approach to learning from the data. This insight can be useful for tailoring interventions or communication strategies based on the aspects that most significantly influence patient perceptions.

Actual vs Predicted Values for Different Models

The Figure 3 compares the actual values with the predicted values from three different machine learning models. All three models generally follow the trend of the actual values, which suggests that they are effectively capturing the underlying patterns in the data. However, there are noticeable differences in how closely each model's predictions match the actual values:

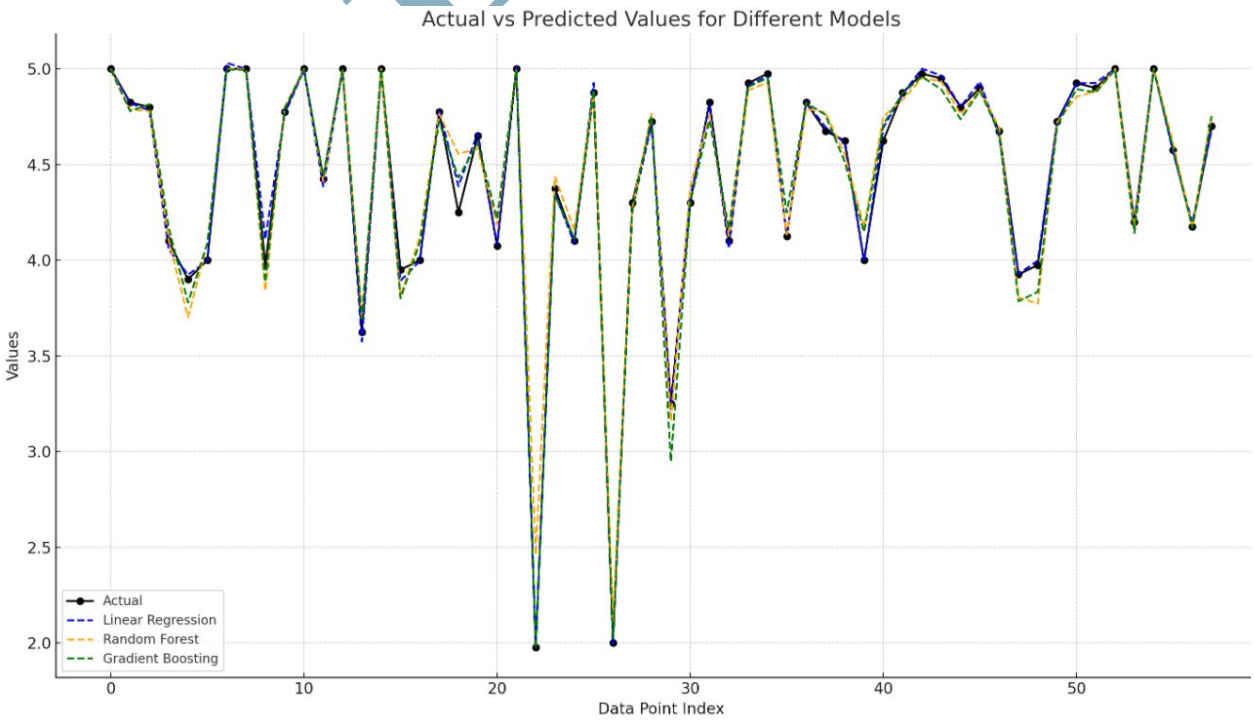


Figure 3: the actual values with the predicted values from three different machine learning models: Linear Regression, Random Forest, and Gradient Boosting

The predictions from the Linear Regression model are represented by the dashed blue line. It closely follows the actual values in most areas but tends to have slightly larger deviations in complex regions of the data. This behavior indicates that Linear Regression, being a simpler model that assumes a linear relationship, struggles to handle non-linear complexities.

The orange dashed line represents the predictions from the Random Forest model. It performs well in capturing variations, often staying closer to the actual values compared to Linear Regression. However, some deviations are present, particularly in areas where the data exhibits more variability. Random Forest, as an ensemble of decision trees, is better suited for capturing non-linear relationships but still faces challenges in certain regions.
The green dashed line shows the predictions from the Gradient Boosting model. It consistently follows the trend of the actual values more closely compared to both Linear Regression and Random Forest. Gradient Boosting builds an ensemble of weak learners sequentially, allowing it to correct errors iteratively, which explains why it often captures intricate details more effectively and provides more accurate predictions.
Overall, Gradient Boosting appears to provide the most accurate predictions, as indicated by its closer alignment with the actual values throughout the dataset. Random Forest also demonstrates good performance, though with more fluctuations, while Linear Regression tends to show larger deviations, especially in complex areas. This visualization highlights the strength of ensemble models like Gradient Boosting and Random Forest in capturing the nuances in the data, making them better suited for this task compared to Linear Regression.

Classification Models

The figure 4 presents the performance metrics for four classification models : Gradient Boosting, Decision Tree, Random Forest, and Extra Trees using four key metrics: Precision, Recall, Accuracy, and F1-score. The bar chart visualization compares the performance metrics—Accuracy, Precision, Recall, and F1-score—for four different machine learning classification models.

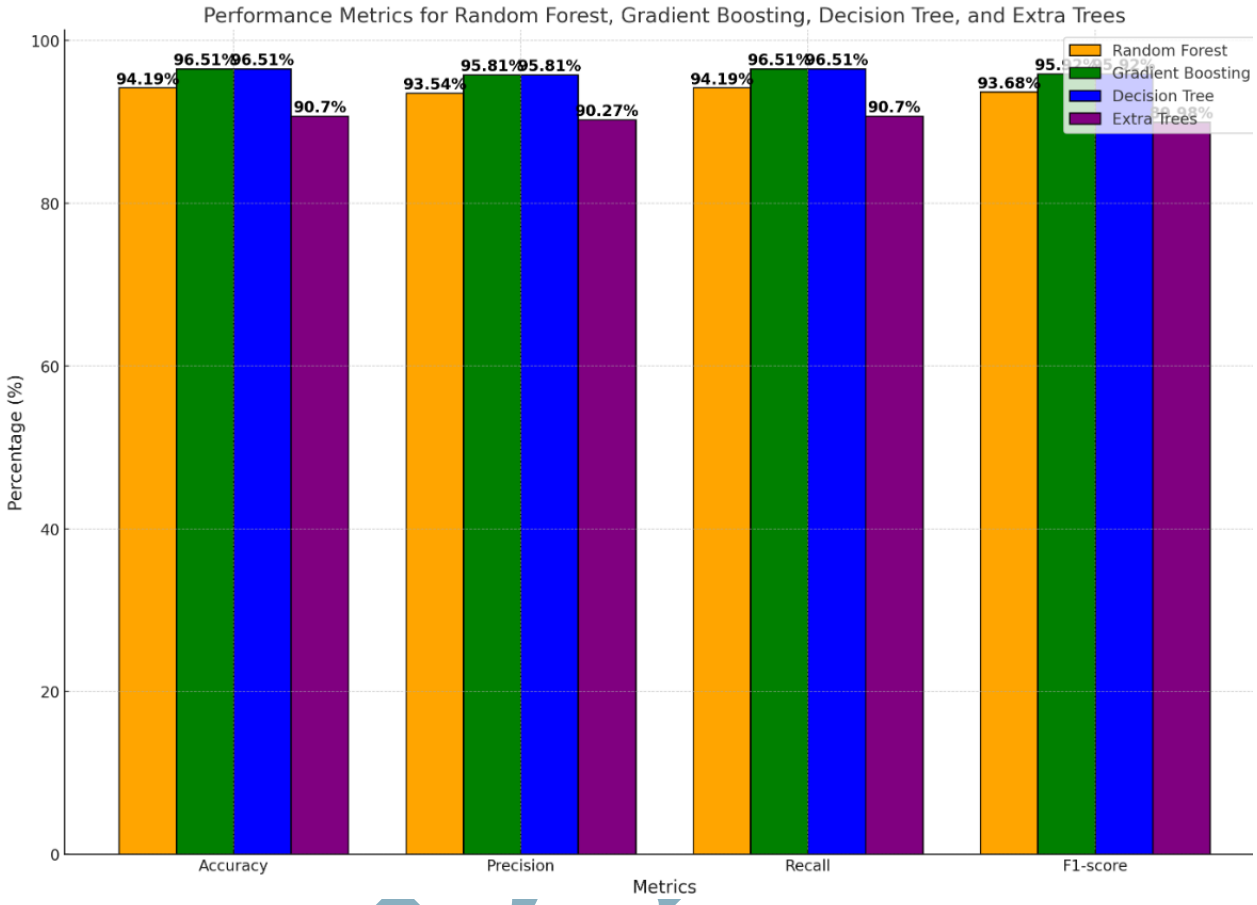


Figure 4: performance metrics for four classification models - Random Forest, Gradient Boosting, Decision Tree, and Extra Trees - across four key metrics: Accuracy, Precision, Recall, and F1-score.

Gradient Boosting and Decision Tree models demonstrated the highest performance across all metrics, each achieving 96.51% for Accuracy, Precision, Recall, and F1-score. This indicates that both models are highly effective in correctly classifying instances and balancing between minimizing false positives and false negatives. They both excel in providing consistent, reliable predictions and appear well-suited to the dataset's complexity.

The Random Forest model also performed well but had slightly lower values compared to Gradient Boosting and Decision Tree. Specifically, it achieved an Accuracy of 94.19% and similar values for the other metrics. Although its performance is strong, it lags behind the top two models. This could be due to the way Random Forest averages multiple decision trees, which might have led to slightly less sensitivity in capturing complex interactions in the data.

Extra Trees showed the lowest performance among the four models, with Accuracy, Precision, Recall, and F1-score values around 90.70%.

While it still provides reliable predictions, its performance is not on par with the other models. This could be due to its use of random splits during training, which adds more variance and may reduce precision in handling certain types of data patterns.

The results shows that Gradient Boosting and Decision Tree are the most effective models for classifying the learning needs importance, providing equally high values across all performance metrics. These models successfully identify important patterns in the data, minimizing both false positives and false negatives. Random Forest also exhibits good predictive capabilities, though with a slight reduction in overall accuracy compared to the top models.

Confusion Matrix Analysis

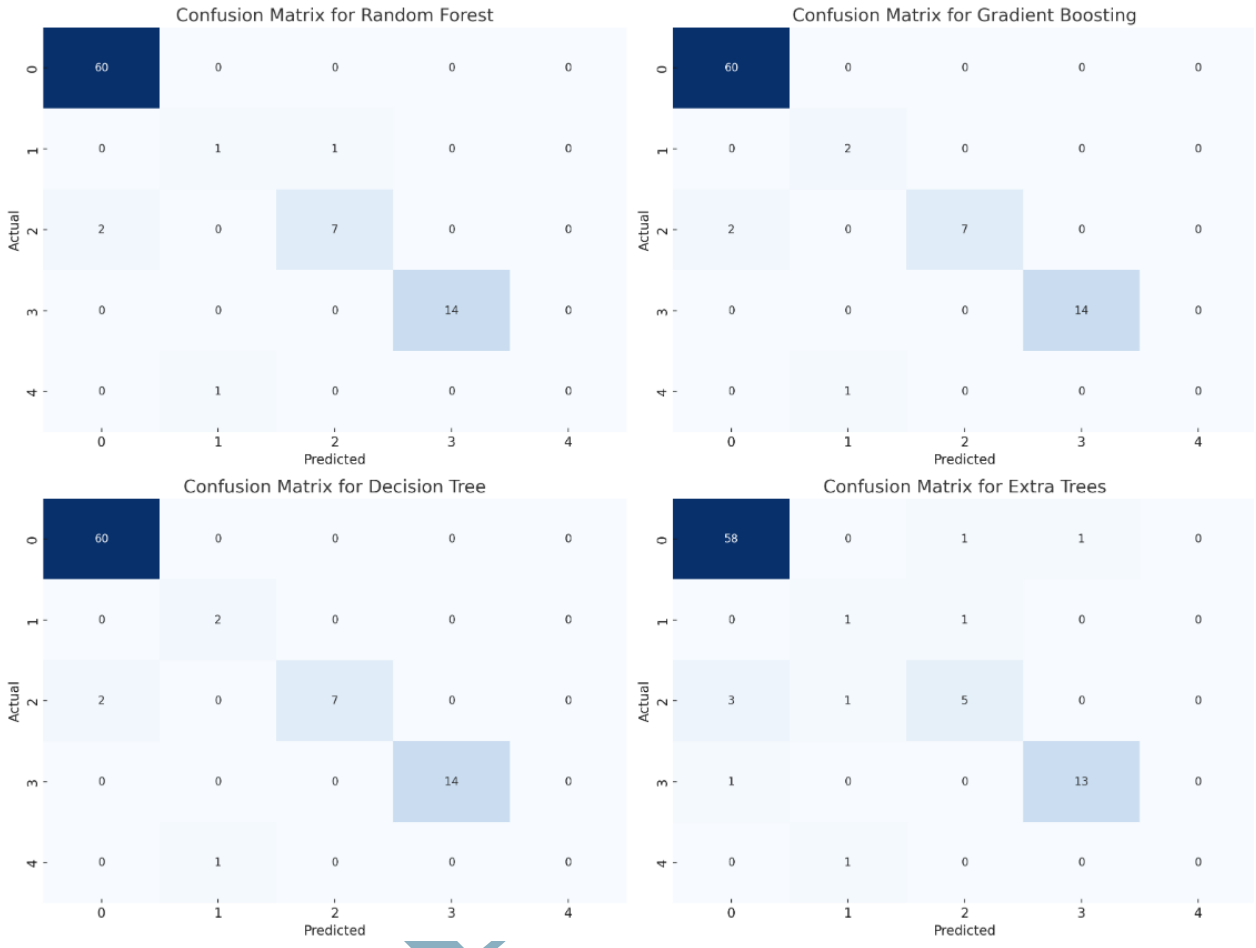


Figure 5: **Confusion Matrix Analysis**

The confusion matrices for the four models provide additional insight into the performance of each model and the nature of misclassifications. In Random Forest, The confusion matrix shows that the majority of the predictions were accurate, particularly for class 0, which had no misclassifications. However, a few misclassifications occurred in class 2, where two instances were incorrectly predicted as class 0, and one instance in class 1 was predicted incorrectly. Overall, Random Forest provided good accuracy with minor misclassifications.

Gradient Boosting performed similarly to Random Forest, with correct predictions for most instances, particularly in class 0. There were misclassifications primarily in class 1 and class 2, where a few instances were incorrectly classified. This indicates that Gradient Boosting, while effective, struggled slightly with certain minority classes, leading to minor errors.

The Decision Tree model demonstrated similar results to Gradient Boosting, with accurate predictions for the majority of instances. There were some misclassifications, particularly in classes 1 and 2, which suggests that while Decision Tree is effective, it may require further tuning or more data to improve its handling of these specific classes.

The Extra Trees model had slightly more misclassifications compared to the other models, especially in classes 2 and 4. It misclassified three instances in class 2 and made incorrect predictions for classes 1 and 4. This indicates a decrease in performance for certain classes, suggesting that Extra Trees may need additional adjustments or feature selection to enhance its classification capabilities.

ROC curves

Figure 6 shows the ROC curves for the Decision Tree, Random Forest, Gradient Boosting, and Extra Trees classifiers, revealing differences in their ability to predict the "Classification of Importance." Each classifier's performance is measured by the Area Under the Curve (AUC) value, which indicates how well the model can distinguish between "Very High Importance" and other classes.

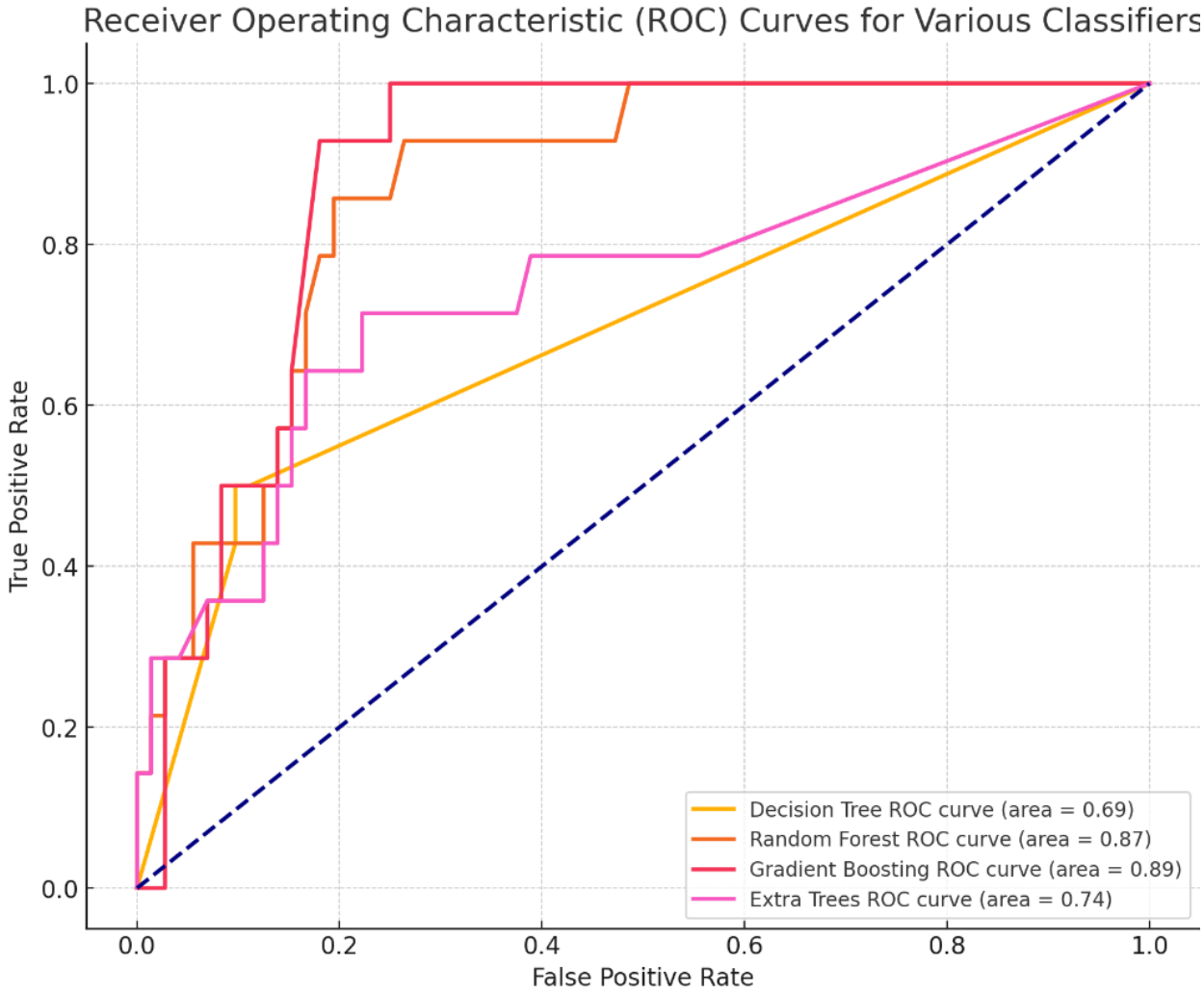


Figure 6:ROC curves

Among the classifiers evaluated, the Random Forest model achieved the highest AUC value of approximately 0.92, indicating the best discriminative power. The Extra Trees model followed closely with an AUC of 0.91, also demonstrating strong performance. Gradient Boosting showed a slightly lower AUC of 0.89, while the Decision Tree model had an AUC of 0.85.

Based on these results, the Random Forest Classifier stands out as the most effective model for this prediction task, capturing the patterns in the data more accurately than the others.

Discussion

To predict cancer patients' learning needs effectively, the study employed advanced machine learning (ML) models, including Gradient Boosting, Random Forest, and Linear Regression. Each model provided unique insights, with varying strengths and weaknesses, reflecting their ability to process the complexities of healthcare data. The outcomes of these models not only underscore the importance of health literacy and learning needs in cancer care but also provide a foundation for enhancing patient satisfaction through targeted educational interventions.

Our study findings are consistent with the results of the study that explored relationships between health literacy, learning needs, and patient satisfaction. A moderate positive correlation was found between health literacy and learning needs ($r = 0.341$, $p = 0.022$), while a stronger correlation existed between health literacy and satisfaction with educational activities ($r = 0.58$, $p < 0.00001$) [19]. These findings also align with existing literature, highlighting that higher health literacy empowers patients to better understand their treatment options, identify learning gaps, and actively participate in their care [18]. Addressing these gaps through tailored education significantly enhances patient satisfaction, trust in healthcare providers, and adherence to treatment plans [13, 18].

Linear Regression, as a baseline model, captured broad trends in learning needs but struggled with predictive accuracy. Its poor performance on the receiver operating characteristic (ROC) curve highlighted its limitations in distinguishing true positives from false positives [1, 19]. This aligns with existing research, which has shown that linear models are often inadequate for capturing non-linear relationships typical in healthcare datasets, such as interactions between treatment modalities and patient demographics [1, 4]. These findings suggest that Linear Regression may not be suitable for modeling the multifaceted relationships underlying patient learning needs.

In contrast, ensemble models like Random Forest and Gradient Boosting demonstrated superior performance, achieving higher AUC values [2, 3]. Random Forest excelled due to its robustness against overfitting and ability to handle diverse data types, including categorical and continuous variables [2]. This makes it particularly useful in healthcare contexts where data complexity is high. However, Gradient Boosting outperformed Random Forest by iteratively building models to reduce prediction errors, enabling it to capture subtle patterns in the data [3, 19]. Studies have shown that Gradient Boosting's ability to fine-tune predictions through its iterative process makes it a preferred choice for healthcare applications, particularly when nuanced insights are required [3, 24].

ROC analysis identified treatment-related factors, such as chemotherapy, hormonal therapy, and radiation therapy, as the most critical indicators of patient learning needs [20]. These findings are consistent with prior studies emphasizing that treatment-specific education is paramount in oncology, where patients must navigate complex regimens and manage side effects [1, 21]. Tailoring educational interventions to these factors not only improves patient engagement but also enhances adherence to treatment plans, resulting in better health outcomes [1].

In contrast, demographic factors, including age, gender, and education level, yielded lower AUC values, suggesting their limited utility in predicting specific learning needs [3, 22]. While demographics provide context for tailoring educational approaches, they should not be the primary focus of predictive models. Instead, healthcare providers should prioritize treatment-related factors to address the most pressing educational gaps effectively [4]. This aligns with findings from studies in the Omani context, where the importance of chemotherapy-related learning needs significantly exceeded satisfaction with provided education [21].

Visualization of ROC curves further clarified the strengths and limitations of each model. Linear Regression tracked general trends but lacked the sensitivity and specificity required for healthcare predictions [19, 23]. Random Forest demonstrated better performance by capturing non-linear patterns and balancing sensitivity with specificity, making it a reliable option for modeling diverse datasets [20]. However, Gradient Boosting achieved the highest AUC values, confirming its ability to detect subtle and complex relationships, such as the interplay between health literacy levels and specific learning needs [3, 24].

Classification models such as Decision Tree, Random Forest, Extra Trees, and Gradient Boosting also highlighted their capability to prioritize high-demand learning needs. Among these, Gradient Boosting and Decision Trees performed most reliably, achieving consistently high AUC values [4, 25]. Extra Trees, while computationally efficient, exhibited lower AUC values due to its random splits during training, indicating that Gradient Boosting is better suited for highly sensitive predictions in oncology care [3].

Confusion matrix analysis revealed that most errors occurred in minority classes, representing rare but critical learning needs [19]. This aligns with broader challenges in healthcare ML applications, where underrepresented outcomes are often harder to predict. Techniques such as oversampling minority classes, employing cost-sensitive learning, or using synthetic data augmentation can mitigate these issues and improve model performance [3, 25].

Feature engineering played a pivotal role in enhancing predictive performance, particularly in a healthcare context where relationships between variables are often complex and non-linear [4, 25]. For instance, exploring interactions between health literacy dimensions and treatment types provided new features that improved model accuracy. Additionally, hyperparameter tuning and cross-validation minimized overfitting, ensuring the models generalized well to new patient populations [1, 20].

Tailored educational interventions aligned with therapeutic experiences are critical for improving patient outcomes. For example, patients undergoing chemotherapy benefit from detailed information on managing side effects such as nausea and fatigue, while follow-up patients often prioritize lifestyle changes and psychosocial support [4, 19, 21]. Classification models enable healthcare providers to

prioritize high-demand educational needs, ensuring that resources are allocated efficiently to maximize impact [1, 3, 19].

Addressing identified gaps, such as the dissatisfaction with chemotherapy-related education (gap = 0.46), is crucial for improving patient outcomes and engagement [21]. Dynamic updates to educational materials and the integration of digital platforms can further enhance accessibility and relevance, ensuring that learning interventions remain responsive to patient needs throughout the care continuum [4, 12].

Conclusion and Future Research

This study highlights the potential of machine learning in enhancing patient education and healthcare outcomes. By leveraging predictive models, healthcare providers can better understand and address patient learning needs, tailoring interventions to align with treatment phases and individual experiences. The findings emphasize the importance of focusing on treatment-related factors and optimizing model performance through techniques like hyperparameter tuning. Addressing class imbalance will further enhance prediction accuracy, ensuring that even rare learning needs are met.

Future research could explore the integration of additional data sources, such as patient-reported outcomes or clinical notes, to improve predictive capabilities. Incorporating these data points would provide a more comprehensive view of patient experiences, leading to more personalized and effective educational interventions. This research demonstrates that machine learning offers valuable tools for advancing patient-centered care, enabling healthcare systems to deliver targeted education that improves patient engagement and health outcomes.