Heart Disease Prediction Using Machine Learning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Student Name  dept. name of organization  (of Affiliation)  name of organization  (of Affiliation)  email address or ORCID  Student Name  dept. name of organization  (of Affiliation)  name of organization  (of Affiliation)  email address or ORCID | Student Name  dept. name of organization  (of Affiliation)  name of organization  (of Affiliation)  email address or ORCID  Student Name  dept. name of organization  (of Affiliation)  name of organization  (of Affiliation)  email address or ORCID | Student Name  dept. name of organization  (of Affiliation)  name of organization  (of Affiliation)  email address or ORCID |  |  |  |

***Abstract*—** **Cardiovascular Disease (CVD) is a growing global health concern, with unhealthy lifestyle choices being significant risk factors. Timely diagnosis is essential to reduce the burden of CVD and its correlation with Coronavirus raises further concern. To address this issue, machine learning models are proposed to predict the presence of cardiovascular illnesses based on health data. Existing methods, namely Random Forest Classifier and Support Vector Machine (SVM), are evaluated for their strengths and weaknesses. However, they have limitations, prompting the proposal of GradientBoostingClassifier (GBC). GBC is an ensemble learning method that sequentially builds weak learners to correct errors and enhance predictive performance. The study compares the performance of GBC, RFC, and SVM. The results show that GBC outperforms the other models in accuracy, recall, and F1 score, making it a promising technique for predicting cardiovascular diseases.**

***Keywords—Support vector machine, Precision, Cardiovascular Disease, Gradient Boosting Classifier*.**

1. INTRODUCTION

The increasing prevalence of Cardiovascular Disease (CVD), characterized by conditions such as ischemic heart diseases and strokes, has become a major global health concern. Unhealthy lifestyle choices, including poor diet, sedentary behavior, tobacco use, and excessive alcohol consumption, are identified as significant risk factors contributing to the development of CVD [10]. The early diagnosis of warning indicators are essential for decreasing the burden of cardiovascular disease and preventing future worsening of one's health. Because there is a possibility of a correlation between Coronavirus and heart disease, there has been increased concern regarding the timely identification and treatment of the condition. In light of the pressing need to solve this matter, there is an urgent requirement for the creation of machine learning models that are both efficient and accurate. These models should be able to determine the chance of acquiring cardiovascular disease based on significant health data and risk factors. The use of machine learning in the field of medicine has already proven its value by making significant contributions to the diagnosis, evaluation and prognosis of a wide range of medical conditions; as a consequence of this, the objective of this research is to propose and carry out an approach that is founded on machine learning, with the goal of predicting the presence of cardiovascular disorders in patients by making use of relevant health data.

II. EXISTING METHODS ALGORITHMS TO SOLVE THE PROBLEM (EXISTING SYSTEM)

# A. Random Forest Classifier

Random Forest is an ensemble learning method that combines a large number of decision trees in order to produce a model that is reliable and accurate for classification tasks [1]. Each individual decision tree that constitutes the Random Forest is built using a random subset of the training data as well as a random subset of the attributes; this unpredictability helps to add to the diversity of the individual trees, which in turn helps to reduce instances of overfitting and improves the model's capacity to generalize.

During the construction of the Random Forest, the algorithm employs a technique called bootstrap aggregation (bagging), which involves sampling the training data with replacement to create different subsets for each tree. Random feature selection is applied at each split point of the tree, considering only a fraction of the total features; this randomness ensures that each tree focuses on different aspects of the data, further increasing the model's diversity.

The decision trees in the Random Forest are grown until specific stopping criteria are met, such as a predefined maximum depth or the minimum number of samples required to split a node. The trees are optimized to minimize information gain, Gini impurity, or entropy, depending on the classification task.

After all of the trees have been formed, they collaborate in order to make predictions regarding the classes of newly acquired data points. The results of the predictions made by each tree are counted as votes, and the category that received the most votes is used to make the ultimate prediction; this helps to ensure that the results are reliable and accurate.

The evaluation process in Random Forest uses what are known as out-of-bag, or OOB, samples [2]. The performance of a particular tree is evaluated based on the results of tests conducted on samples that were not included in its training set; this is due to the fact that each tree is trained on a unique subset of the total training data. Estimating the performance of the model does not require the use of a separate validation set as long as it is done by taking the accuracy of each tree on the OOB samples and average them; these features, combined with the simplicity of implementation and ability to handle large datasets, make Random Forest a popular choice for various classification tasks in different domains.

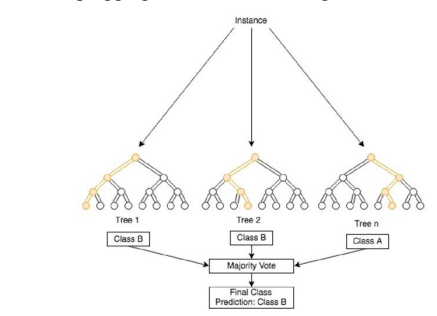


Fig. 1. Random forest classifier

# B. Support Vector Machine

The Support Vector Machine, is a robust supervised machine learning technique that can be utilized for classification and regression work [3]. Figure 1 illustrates the fundamental concept underlying SVM, which is to locate the ideal hyperplane that most effectively differentiates between the two classes in the feature space. The hyperplane serves as a decision boundary by increasing the distance between the two classes' data points that are the most similar to one another. Because these data points are considered to be "support vectors," the machine in question is referred to as a "Support Vector Machine."

The data is represented in SVM as feature vectors in a multi-dimensional space and each feature corresponds to a particular attribute or characteristic of the data; when the data is linearly separable, which means that the two classes can be completely separated by a hyperplane, SVM performs at its highest potential. In the event that the data cannot be segmented linearly, the SVM will employ the kernel trick in order to map the data into a space with a higher dimension; In the future, linear separation of the data will become feasible as a result of this.

This can be accomplished by finding the hyperplane that maximizes this margin; first, the support vectors, which are the essential data points that lie on the margin or those that are incorrectly classified, are determined, and then using that information, this hyperplane is found. The primary objective of the SVM classification method is to locate the hyperplane that maximizes the margin between the hyperplane and the data points of each class that are closest to the hyperplane. The critical data points that fall on the margin or those that are improperly categorised are what we refer to as support vectors. The data points known as support vectors are extremely important and are located on the margins; this endeavor is represented by SVM as an optimization problem, and its goal is to reduce as much as possible the weight vector that represents the hyperplane; all the while guaranteeing that each and every data point is correctly classified or lies inside the margin.

SVM makes use of a variety of kernel functions, including polynomial kernels, radial basis function (RBF) kernels, and sigmoid kernels, in order to manage non-linearly separable data [4]. These kernels take the original feature space and change it into a higher-dimensional space, which then enables linear separation in the space that has been transformed. The support vector machine (SVM) utilizes a regularization parameter denoted by the letter C to control the balance between the two goals of maximizing the margin and minimizing classification errors. A higher C value places more of an emphasis on correct classification whereas a lower C value places more of an emphasis on having a greater margin, which may lead to some misclassifications. When the best hyperplane has been found, it is simple to classify fresh data points by determining which side of the hyperplane they fall on. The capability of SVM can handle both linear and non-linear data, in addition to its capacity for regularization, makes it an adaptable and powerful method for a wide variety of classification applications.

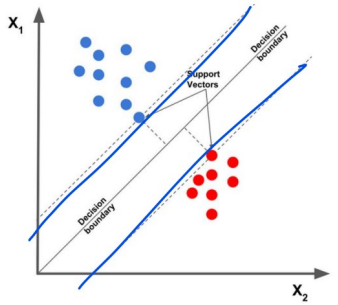


Fig. 2. Support vector machine

III. GRADIENT BOOSTING CLASSIFIER AS A PROPOSED ALGORITHM

The Gradient Boosting Classifier is an ensemble learning method that creates a powerful predictive model by building numerous weak learners in a sequential fashion [5]. It does this by combining the predictions made by these less capable learners in order to increase accuracy and generalization performance overall; it is the goal of the Gradient Boosting algorithm to improve upon the accuracy of the predictions provided by earlier models by improving upon the errors that were made by those models.

The Gradient Boosting Classifier begins by initializing the first weak learner, typically a shallow decision tree, which generates initial predictions [6]. As these initial predictions are prone to errors, especially in complex data scenarios, the subsequent steps aim to improve the overall model performance.

In order for the algorithm to accomplish this goal, a loss function is defined. This function measures the amount of deviation that exists between the values that are predicted and the actual target values that are present in the training data [11]. The nature of the problem should guide the specific decision about the loss function; for example, the mean squared error should be used for regression tasks, whereas log loss should be used for binary classification.

As the method attempts to minimize the loss function, the concept of gradient descent is brought into play. It does so by computing the negative gradient of the loss function with regard to the predictions that the current model has made; this gradient outlines the path that needs to be taken in order to make the necessary adjustments to the model's predictions in order to reduce the amount of loss.

This procedure seeks to rectify the errors caused by the prior model and concentrate on data points that were previously difficult to reliably forecast. With the gradient information in hand, a new weak learner, which is often another decision tree, is trained to fit the negative gradient of the loss function.

After that, the predictions of the new model are integrated with those of the earlier models. The contribution of each model is then weighted, taking into consideration parameters such as the learning rate, which influences the step size during gradient descent and the model's influence on the final prediction.

The iterative process continues, repeating steps 3 to 5, to construct additional weak learners. Each new model further corrects errors from previous iterations, gradually enhancing the overall predictive performance.

After reaching a predetermined threshold for the percentage of ineffective students, the conclusive forecast is arrived at by adding up the results of each of the models. In classification problems, the final forecast is obtained by employing majority voting, but in regression tasks, it is derived as the weighted average of the guesses. In both cases, the prediction is based on the best available evidence. This culminates in a powerful Gradient Boosting Classifier capable of handling various types of data and producing robust predictions.

# A. Why propose Gradient Boosting Classifier

We proposed Gradient Boosting Classifier with aim of increasing the model performance. This is because, Gradient Boosting performs well with noisy data and can adapt to such noise through its boosting iterations [12]. Gradient Boosting Classifier is also an ensemble method that builds multiple weak learners (decision trees) sequentially, with each one correcting the errors of the previous one; this leads to improved predictive performance, as it combines the strengths of individual models.

# B. Drawbacks in the previous work

The model checks for class imbalance but concludes that no resampling is needed based on a 9:11 class ratio. However, the severity of class imbalance and its impact on the model's performance should be thoroughly analyzed. The implemented models had above average accuracy but not as good as expected.

IV. QUALITATIVE AND QUANTITATIVE COMPARATIVE ANALYSIS

Support Vector Machine (SVM) is a powerful classification algorithm with several strengths. It performs effectively in high-dimensional spaces, making it suitable for complex datasets [7]. SVM is versatile and can handle non-linear data through various kernel functions. It also works well when the number of features exceeds the number of samples; however, SVM has some weaknesses. One of its drawbacks is its computational expense, especially for large datasets. Achieving optimal performance requires careful tuning of hyperparameters and the choice of the kernel function and regularization parameter can significantly impact the results.

The Random Forest Classifier is a popular classification method with its own set of strengths. One of the main advantages of random forests is their robustness to overfitting, attributed to the ensemble of multiple decision trees [8]. They handle both categorical and numerical features with minimal preprocessing and can effectively handle large datasets with high dimensionality. Despite these advantages, random forests are often difficult to interpret compared to single decision trees; they may also be prone to overfitting if the number of trees in the ensemble is excessively large. Random forests might be less effective at capturing linear relationships in the data.

The Gradient Boosting Classifier is a sequential ensemble model that addresses the weaknesses of weak learners by improving their performance at each iteration [9]. Similar to random forests, gradient boosting handles various data types well and requires minimal data preprocessing. It can handle missing data effectively without the need for imputation. However, gradient boosting is more sensitive to overfitting compared to random forests, necessitating careful hyperparameter tuning to achieve the best results. This method is computationally more expensive than random forests due to its sequential nature.

V. RESULTS OF ITS QUANTITATIVE AND QUALITATIVE COMPARISON

# A. Results of the Predicted Model before Removing Extreme Outliers

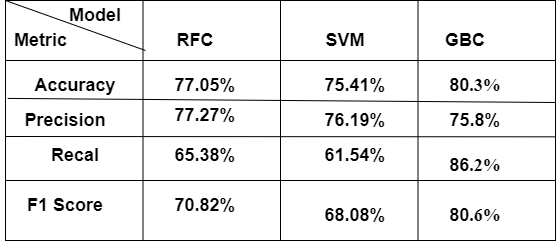


Fig. 3. Results of the predicted model before removing extreme outliers

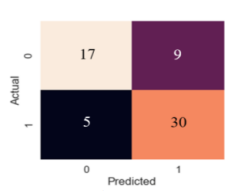


Fig. 4. RFC confusion matrix

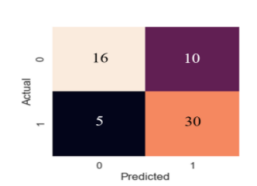


Fig. 5. SVM confusion matrix

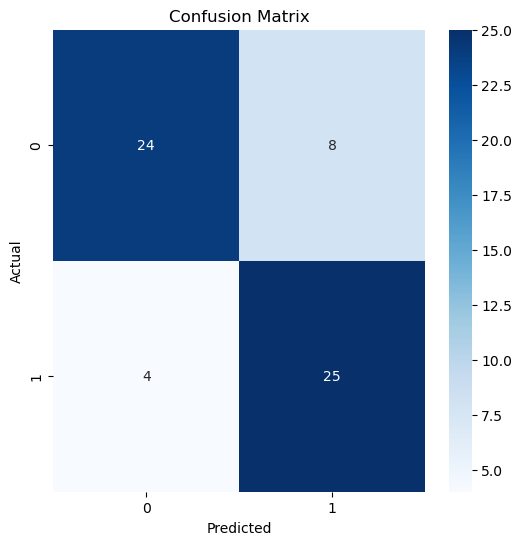


Fig. 6. GBC confusion matrix

RFC achieved an accuracy of 77.05%, SVM achieved 75.41%, and GBC performed the best with an accuracy of 80.30%. The Gradient Boosting Classifier (GBC) demonstrated the highest accuracy, indicating its superior ability to predict the presence of cardiovascular diseases compared to the other two models.

RFC achieved the highest precision of 77.27%, followed closely by SVM with 76.19%, and GBC had a precision of 75.80%. The Random Forest Classifier (RFC) displayed the best precision, meaning that when it predicted a patient to have a cardiovascular disease, it was correct 77.27% of the time.

GBC achieved the highest recall of 86.20%, outperforming RFC with 65.38% and SVM with 61.54%. The Gradient Boosting Classifier (GBC) identified 86.20% of the actual cases of cardiovascular disease in the dataset, indicating its ability to effectively capture positive instances.

GBC achieved the highest F1 score of 80.60%, while RFC scored 70.82%, and SVM scored 68.08%. The F1 score, which balances precision and recall, further highlights the superior performance of the Gradient Boosting Classifier (GBC) in terms of overall predictive capability.

The Gradient Boosting Classifier demonstrated the best performance across accuracy, recall and F1 score when compared to the Random Forest Classifier and Support Vector Machine.

# B. Results of the Predicted Model after Removing Extreme Outliers

To remove extreme outliers, the Interquartile Range (IQR) method was employed. The IQR was calculated for each column by determining the 25th percentile (Q1) and the 75th percentile (Q3) of the data. The IQR was then obtained by subtracting Q1 from Q3. The threshold for extreme outliers was defined. Any data points that fell below Q1 - 3 \* IQR or above Q3 + 3 \* IQR were considered extreme outliers. To apply this filtering, a new DataFrame called "df\_filtered" was created, which only included the rows that met the condition of being within the defined thresholds for all columns; this process helps in cleaning the data and ensures that extreme outliers do not unduly influence statistical analyses or machine learning models.

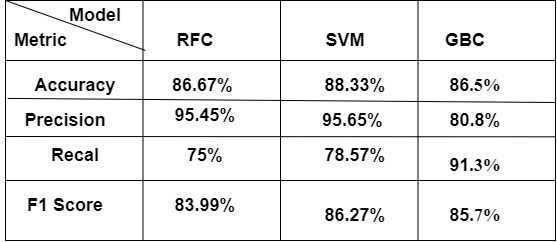


Fig. 7. Results of the predicted model after removing extreme outliers

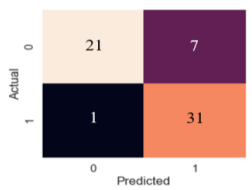


Fig. 8. Random forest confusion matrix

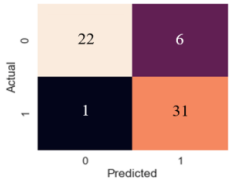


Fig. 9. SVM confusion matrix

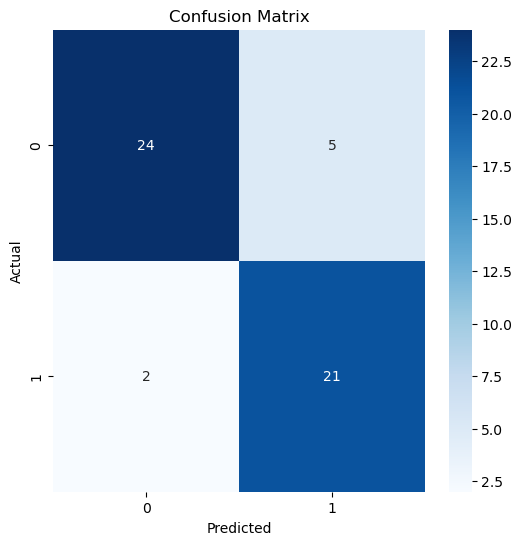


Fig. 10. GBC confusion matrix

The Gradient Boosting Classifier (GBC) achieved the highest performance, reaching an accuracy of 80.30%. It outperformed both the Random Forest Classifier (RFC), which achieved an accuracy of 77.05% and the Support Vector Machine (SVM), which achieved an accuracy of 75.41%; this indicates that GBC is more effective in correctly classifying patients with and without cardiovascular diseases compared to the other two models.

The Random Forest Classifier (RFC) obtained the highest score of 77.27%. When RFC predicted a patient to have a cardiovascular disease, it was correct 77.27% of the time. SVM achieved a precision of 76.19%, while GBC achieved a precision of 75.80%. Although RFC exhibited the highest precision, GBC came close and demonstrated strong precision performance as well.

The Gradient Boosting Classifier (GBC) excelled by achieving the highest recall of 86.20%. This implies that GBC effectively identified a substantial portion (86.20%) of the actual cases of cardiovascular disease in the dataset. On the other hand, RFC attained a recall of 65.38%, and SVM achieved a recall of 61.54%, both of which were lower than GBC's performance.

The Gradient Boosting Classifier (GBC) achieved the highest score at 80.60%. This signifies that GBC strikes a better balance between precision and recall compared to the other two models, RFC (F1 score: 70.82%) and SVM (F1 score: 68.08%).

The Gradient Boosting Classifier (GBC) demonstrated superior performance across accuracy, recall, and F1 score when compared to both the Random Forest Classifier (RFC) and the Support Vector Machine (SVM). However, RFC exhibited the highest precision among the three models.

# C. Results of the Predicted Model after Applying Feature Selection

Feature selection was performed using Recursive Feature Elimination with Cross-Validation (RFECV); the dataset was first split into features (X) and the target variable (y), where the target variable represented whether the patient had cardiovascular disease or not. The Gradient Boosting Classifier was chosen as the estimator for feature selection; the RFECV estimator was initialized with this classifier and the Stratified KFold cross-validation strategy, which is a cross-validation method that maintains the class distribution of the target variable during the splitting process.

The RFECV was then fit on the data, and during this process, it recursively eliminated features and evaluated the performance of the estimator using cross-validation; it started with all features and iteratively removed the least important feature, evaluating the performance after each elimination. The process continued until the optimal number of features was determined, which is the number of features that resulted in the best estimator performance. The selected features were:



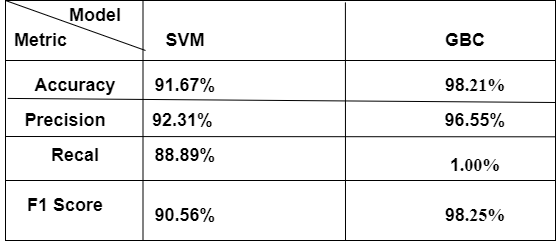
****

Fig. 11. Results of the predicted model after applying feature selection

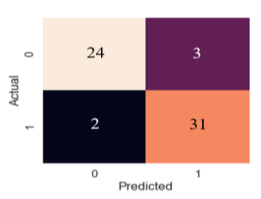
****

Fig. 12. SVM confusion matrix

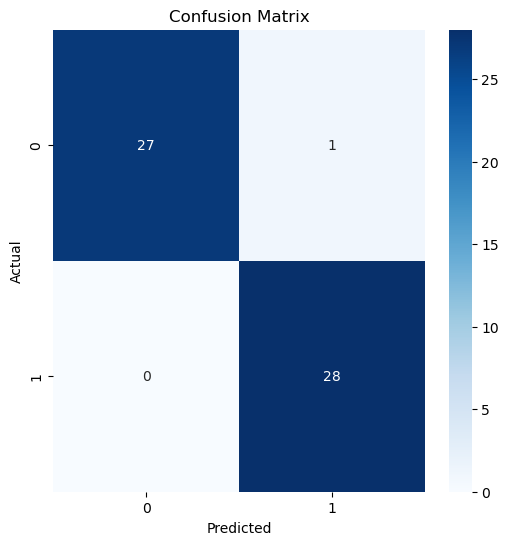


Fig. 13. GBC confusion matrix

The Gradient Boosting Classifier (GBC) exhibited superior accuracy with a remarkable score of 98.21%. On the other hand, the Support Vector Machine (SVM) also demonstrated good accuracy, but comparatively lower than GBC, achieving an accuracy of 91.67%.

The Gradient Boosting Classifier (GBC) outperformed the Support Vector Machine (SVM) in precision, achieving a score of 96.55%. SVM, while still performing well, achieved a precision of 92.31%.

Remarkably, the Gradient Boosting Classifier (GBC) achieved a perfect recall score of 1.00%, implying that it identified all the actual positive cases of cardiovascular disease in the dataset. In contrast, the Support Vector Machine (SVM) achieved a recall of 88.89%, which, though lower than GBC, is still indicative of its ability to identify a substantial portion of actual positive cases.

The F1 score, which balances precision and recall, was impressively high for both models; the Gradient Boosting Classifier (GBC) achieved a score of 98.25%, while the Support Vector Machine (SVM) attained a slightly lower score of 90.56%.

The Gradient Boosting Classifier (GBC) significantly outperformed the Support Vector Machine (SVM) in terms of accuracy, precision, recall and F1 score.

VI. ADVANTAGES AND LIMITATIONS OF THE NEW METHOD/ALGORITHM

# A. Advantages

The Gradient Boosting Classifier (GBC) offers several advantages for predicting the presence of cardiovascular diseases in patients based on relevant health data. Notably, it demonstrated remarkable accuracy with a score of 98.21%, making it a reliable choice for medical diagnosis and decision-making. The high precision of 96.55% is also noteworthy, indicating that when GBC predicts a patient to have cardiovascular disease, it is correct 96.55% of the time. This precision is crucial in medical applications where false positive predictions could lead to unnecessary medical interventions or patient anxiety.

One of the most striking advantages of GBC is its perfect recall score of 1.00%, implying that the model correctly identified all actual positive cases of cardiovascular disease in the dataset. This high recall is highly desirable in medical diagnosis to ensure that no positive cases are missed, reducing the risk of overlooking patients who require medical attention. Moreover, GBC achieved an impressive F1 score of 98.25%, which strikes a balance between precision and recall. This indicates that the model can simultaneously achieve high precision and recall, making it suitable for tasks where both aspects are crucial.

# B. Disadvantages

However, GBC does have some limitations to consider. First, it is computationally expensive due to the sequential combination of weak learners, making it more time-consuming and resource-intensive, especially for large datasets. Additionally, the model's performance heavily relies on the tuning of its hyperparameters, necessitating expert knowledge and effort to achieve optimal results. GBC's sensitivity to outliers is another concern, which could lead to suboptimal performance. Preprocessing steps or outlier handling techniques may be necessary to address this issue. Lastly, GBC's ensemble nature with decision trees makes it less interpretable compared to simpler models like linear regression, which could be challenging for understanding its decision-making process in critical applications where transparency is essential.

VII. APPLICATIONS OF GRADIENT BOOSTING CLASSIFIER

The Gradient Boosting Classifier is a powerful machine learning algorithm that finds applications in various fields due to its high predictive performance and ability to handle complex datasets; some of the possible applications of the Gradient Boosting Classifier include:

* Healthcare and Medical Diagnosis: GBC can be used for medical diagnosis, such as predicting the presence of diseases, like cardiovascular diseases, cancer, diabetes and more, based on patient health data; its high accuracy and recall make it well-suited for identifying critical medical conditions.
* Finance and Risk Management: GBC can be employed in the financial sector for credit risk assessment, fraud detection, and stock market prediction; its ability to handle large datasets and capture complex relationships makes it valuable in assessing financial risks and making data-driven investment decisions.
* Natural Language Processing (NLP): GBC can be applied to various NLP tasks, including sentiment analysis, spam detection, text categorization and document classification; its ability to handle textual data and nonlinear patterns makes it effective in processing and analyzing unstructured text.
* Customer Relationship Management (CRM): GBC can be utilized for customer churn prediction, customer segmentation, and recommendation systems in CRM applications; its ability to handle large feature spaces and nonlinear relationships aids in understanding customer behavior and improving customer engagement.

REFERENCES

1. Naghibi, S.A., Pourghasemi, H.R. & Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess 188, 44 (2016). <https://doi.org/10.1007/s10661-015-5049-6>
2. SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information RetrievalAugust 2015Pages 353–362https://doi.org/10.1145/2766462.2767747
3. M. Somvanshi, P. Chavan, S. Tambade and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, India, 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860040.
4. Pardo, Matteo, and Giorgio Sberveglieri. "Classification of electronic nose data with support vector machines." Sensors and Actuators B: Chemical 107.2 (2005): 730-737. <https://doi.org/10.1016/j.snb.2004.12.005>
5. Feng, De-Cheng, et al. "Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements." Engineering Structures 235 (2021): 111979. <https://doi.org/10.1016/j.engstruct.2021.111979>
6. F. Li, Y. Zuo, H. Lin and J. Wu, "BoostXML: Gradient Boosting for Extreme Multilabel Text Classification With Tail Labels," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3285294.
7. Al-Thanoon, Niam Abdulmunim, Omar Saber Qasim, and Zakariya Yahya Algamal. "A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics." Chemometrics and Intelligent Laboratory Systems 184 (2019): 142-152.https://doi.org/10.1016/j.chemolab.2018.12.003
8. Navega, D., Coelho, C., Vicente, R. et al. AncesTrees: ancestry estimation with randomized decision trees. Int J Legal Med 129, 1145–1153 (2015). <https://doi.org/10.1007/s00414-014-1050-9>
9. Feng, De-Cheng, et al. "Implementing ensemble learning methods to predict the shear strength of RC deep beams with/without web reinforcements." Engineering Structures 235 (2021): 111979. <https://doi.org/10.1016/j.eswa.2018.08.040>
10. Habib, Samira Humaira, and Soma Saha. "Burden of non-communicable disease: global overview." Diabetes & Metabolic Syndrome: Clinical Research & Reviews 4.1 (2010): 41-47. <https://doi.org/10.1016/j.dsx.2008.04.005>
11. Owolabi, T.O., Akande, K.O. & Olatunji, S.O. Estimation of Superconducting Transition Temperature T C for Superconductors of the Doped MgB2 System from the Crystal Lattice Parameters Using Support Vector Regression. J Supercond Nov Magn 28, 75–81 (2015). <https://doi.org/10.1007/s10948-014-2891-7>
12. F. Li et al., "A Light Gradient Boosting Machine for Remainning Useful Life Estimation of Aircraft Engines," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 2018, pp. 3562-3567, doi: 10.1109/ITSC.2018.8569801.