The Application of Varied Machine Learning Algorithms for Obesity Classification

Sri Laya Nalmelwar, Sindhu Buggana, Vaishnavi Nadipelly , Naga Durga Jai Rathan Ghanta

Department of Computer Science, Stevens Institute of Technology (SIT), Hoboken, New Jersey.

Abstract

Obesity is a serious global health issue that leads to both physical and mental health issues. Since obesity is becoming more and more common, more research is required to determine what factors contribute to obesity and how to forecast the condition's occurrence based on these factors. Research has been performed by using traditional machine learning algorithms providing statistical metrices such as accuracy to establish its efficiency and effectiveness. These algorithms include- Naives Bayes, Random Forest, K-Nearest Neighbour, Support Vector Machine, Artificial Neural Network, Logistic regression, Gradient Boost and Ada Boost. The obvious question raised by these implementations is whether ensemble machine learning algorithms, when used for Obesity classification, could perform better individual machine learning algorithms. The focus of this study is indicated by this question. This study used statistical metrics to compare and evaluate eight machine learning algorithms for classifying obesity. The dataset collected from Kaggle was utilized for training and testing the models. The model was validated utilizing four statistical metrices- accuracy, precision, recall and F1 score.

Introduction

Obesity has become a major global health concern that poses serious risks to people's physical and mental health. Given the rising incidence of obesity, it is critical to comprehend the intricate interactions between these variables and create predictive models that are accurate. This research explores the estimation of obesity levels using dietary patterns and physical health indicators, with a particular focus on individuals from Mexico, Peru, and Colombia.

The dataset being examined is made up of 2111 records and 17 attributes, all carefully selected to make it easier to investigate obesity trends in the previously mentioned areas. A class variable called NObesity (Obesity Level) is attached to every record, allowing for accurate categorization into groups like Insufficient Weight, Normal Weight, Overweight Levels I and II, and Obesity Types I, II, and III.

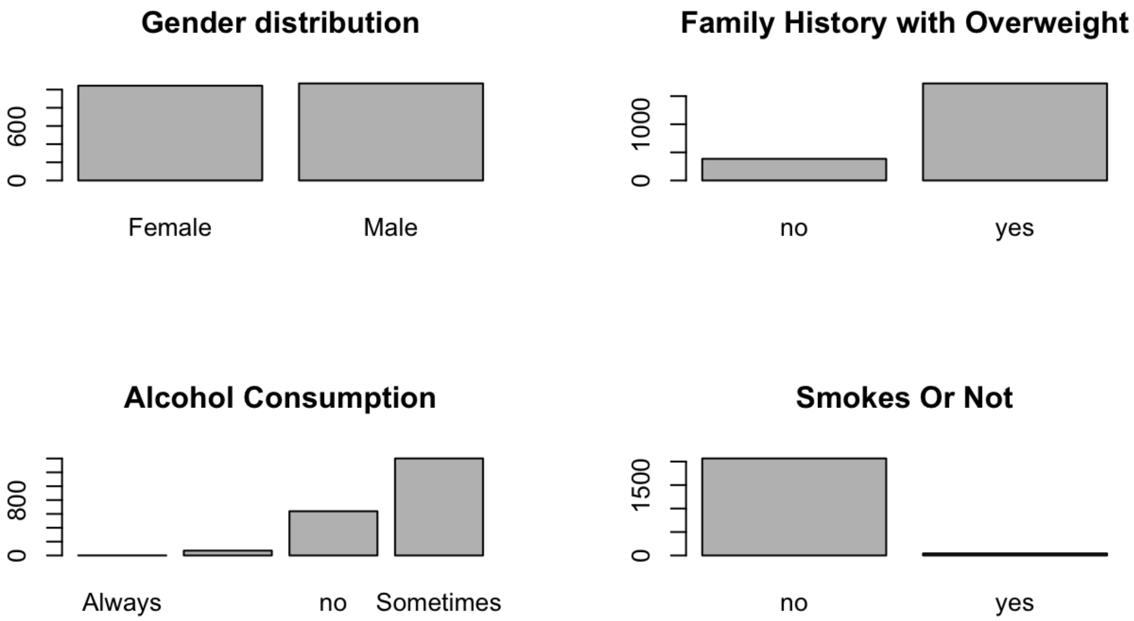
Using cutting edge machine learning techniques, precise analysis and prediction have been the goal. Weka is a powerful tool that was used to synthesize a large portion of the dataset (77%) using the SMOTE (Synthetic Minority Over-sampling Technique) filter. By addressing potential class imbalance issues and ensuring a robust representation of diverse scenarios, this strategic approach improves the reliability of predictive models. Moreover, an additional twenty-three percent of the dataset was obtained directly from users via a specialized online platform, which enhances the dataset's genuineness and applicability to actual situations. By utilizing machine learning techniques like ensemble methods, decision trees, and neural networks, we hope to identify the complex factors that contribute to obesity and create predictive models that can reliably categorize people into different obesity categories.
Our goal is to use the knowledge gained from this dataset to better understand the complex factors that lead to obesity and to help develop targeted interventions and predictive models that lessen the negative effects of obesity on public health.

Research Machine learning modeling

This study compares hybrid and conventional machine learning algorithms for Obesity prediction to find out which provides superior statistical values in terms of accuracy, precision, recall and F1 score using machine learning architecture. The architecture consists of several components, including a data custodian that stores the downloaded data in online storage. Exploratory data analysis ensures that the dataset is properly pre-processed and cleaned before machine learning training. The machine learning algorithms independently use the machine learning data for training, while the model validation and comparative analysis modules test the effectiveness of the model and compare the validation results with existing models. A sequential methodological strategy (SMS) was adopted in the study. A dataset obtained from the online archive Kaggle was used as the machine learning data.

Exploratory Data Analysis

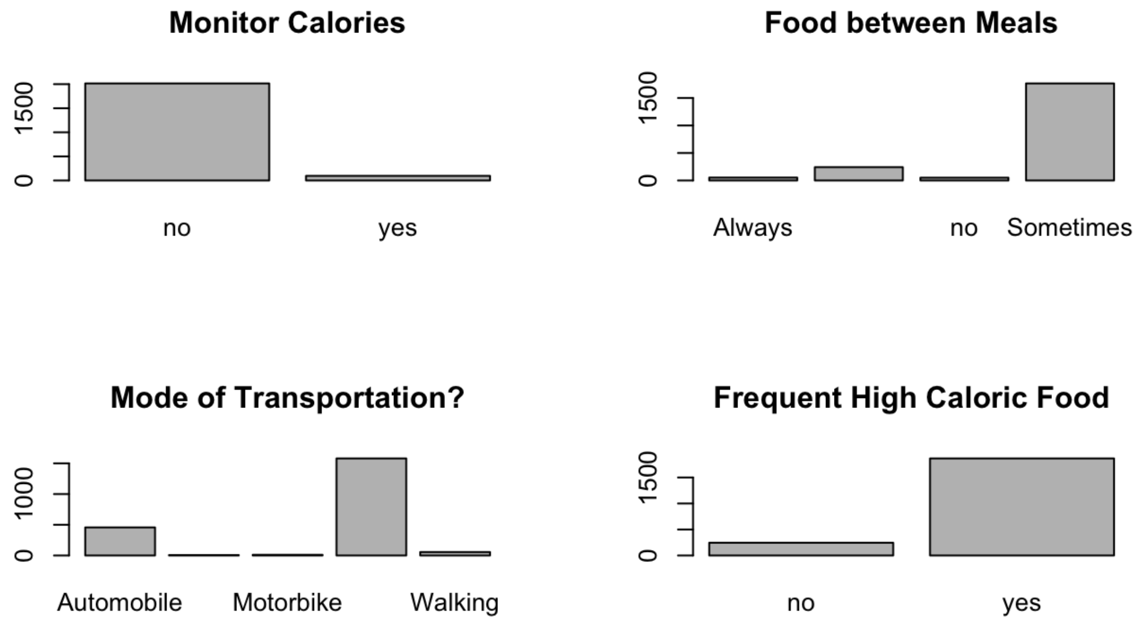
The dataset comprises various features such as gender, age, height, weight, and lifestyle factors like dietary habits, physical activity, and technology usage. With a mix of categorical and continuous variables, it offers insights into individuals' behaviors and health indicators. Notably, a considerable portion of individuals report frequent consumption of high-caloric food and a lack of calorie monitoring. Additionally, a significant proportion has a family history of being overweight.

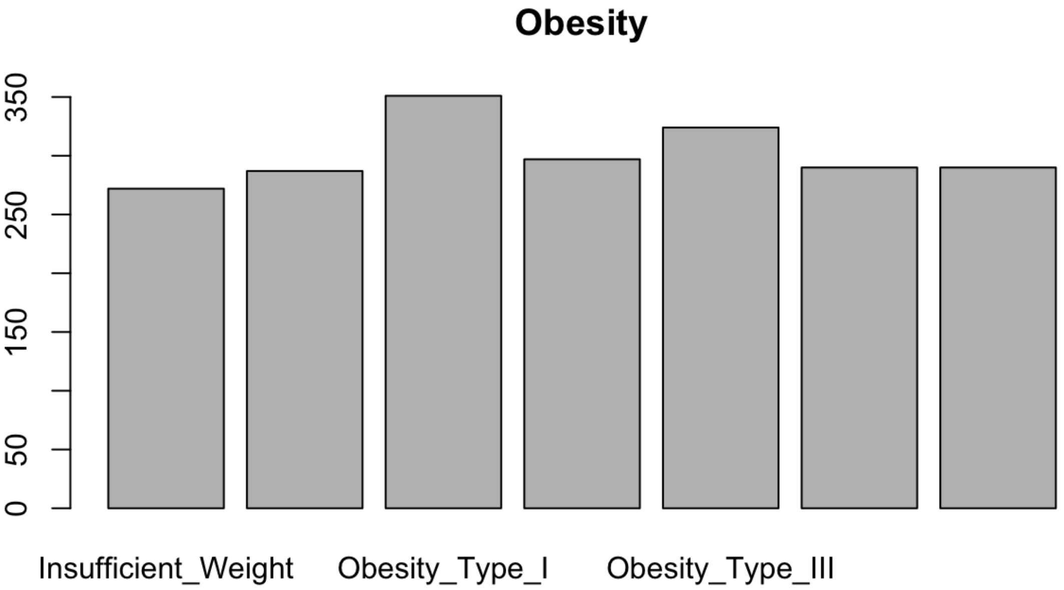

Gender distribution

Family History with Overweight

Alcohol Consumption

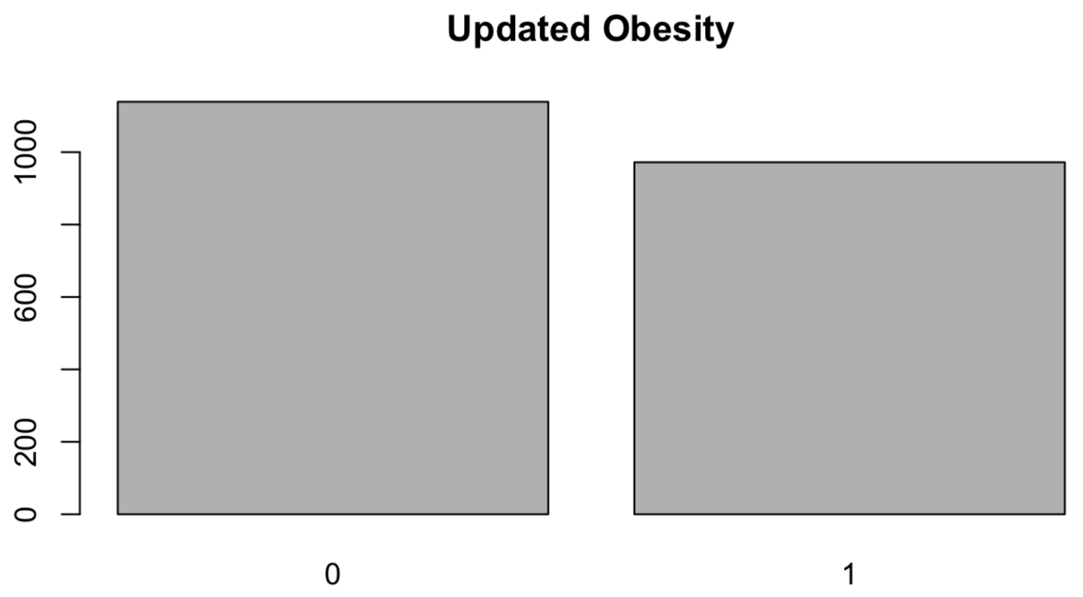
Smokes Or Not





Feature Engineering

After performing feature engineering and converting the Obesity target variable to a binary variable where all obesity levels are changed to "yes" and other values are changed to "no", the dataset now facilitates binary classification, simplifying the prediction task. This transformation enables the development of machine learning models to predict whether an individual is obese or not based on various features like gender, age, lifestyle factors, and health indicators. By framing the problem as a binary classification task, the focus shifts to identifying significant predictors of obesity, thereby aiding in targeted interventions and preventive healthcare strategies.



Implementation of machine learning algorithms

The eight machine learning algorithms codes were implemented utilizing R codes. These algorithms were structured modularly utilizing the dataset obtained from Kaggle. The dataset comprises of 253681 samples with 17 attributes and 2111 records. The dataset was partitioned into 70:30. The implementation of the eight machine learning algorithms was enacted utilizing 70% of the EDA dataset while the model validation was enacted utilizing 30% of the EDA dataset. Below are the machine learning algorithms implemented and the results.

K-Nearest Neighbors (KNN):

KNN is a simple and intuitive classification algorithm that classifies data points based on the majority class among their k-nearest neighbors in the feature space.

In our report, KNN achieved an accuracy of 0.8737, precision of 0.8138, recall of 0.8768, and F1 score of 0.8138. KNN tends to perform well when the decision boundary is not highly nonlinear and when the dataset is not too large.

The accuracy of 0.8737 indicates that KNN correctly classified approximately 87.37% of instances. The precision of 0.8138 suggests that KNN correctly identified around 81.38% of actual positive instances among all instances it classified as positive. The recall of 0.8768 indicates that KNN successfully captured around 87.68% of all actual positive instances. The F1 score of 0.8138 balances both precision and recall, providing a single measure of KNN's performance.

Logistic Regression:

Logistic Regression is a linear classification algorithm that models the probability of a binary outcome using a logistic function.

In our report, Logistic Regression achieved an accuracy of 0.7695, precision of 0.6809, recall of 0.7653, and F1 score of 0.6809.

Logistic Regression is suitable for binary classification tasks and performs well when the relationship between the features and the target variable is approximately linear.

The accuracy of 0.7695 indicates that Logistic Regression correctly classified approximately 76.95% of instances. The precision of 0.6809 suggests that Logistic Regression correctly identified around 68.09% of actual positive instances among all instances it classified as positive. The recall of 0.7653 indicates that Logistic Regression successfully captured around 76.53% of all actual positive instances. The F1 score of 0.6809 provides a balanced measure of Logistic Regression's precision and recall.

Random Forest:

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve classification accuracy.

In our report, Random Forest achieved an accuracy of 0.9251, precision of 0.9096, recall of 0.9306, and F1 score of 0.9096. Random Forest is robust to overfitting and performs well with complex, nonlinear relationships between features and the target variable.

The accuracy of 0.9251 indicates that Random Forest correctly classified approximately 92.51% of instances. The precision of 0.9096 suggests that Random Forest correctly identified around 90.96% of actual positive instances among all instances it classified as positive. The recall of 0.9306 indicates that Random Forest successfully captured around 93.06% of all actual positive instances. The F1 score of 0.9096 provides a balanced measure of Random Forest's precision and recall.

Ada Boost:

Ada Boost is an ensemble learning algorithm that combines multiple weak learners (e.g., decision trees) to create a strong classifier.

In our report, Ada Boost achieved an accuracy of 0.9075, precision of 0.8909, recall of 0.9141, and F1 score of 0.8909.

Ada Boost is effective in improving the performance of weak learners by assigning higher weights to misclassified instances in subsequent iterations.

The accuracy of 0.9075 indicates that Ada Boost correctly classified approximately 90.75% of instances. The precision of 0.8909 suggests that Ada Boost correctly identified around 89.09% of actual positive instances among all instances it classified as positive. The recall of 0.9141 indicates that Ada Boost successfully captured around 91.41% of all actual positive instances. The F1 score of 0.8909 provides a balanced measure of Ada Boost's precision and recall.

Naive Bayes:

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features.

In our report, Naive Bayes achieved an accuracy of 0.9075, precision of 0.8909, recall of 0.9141, and F1 score of 0.8909. Naive Bayes is simple, fast, and performs well with high-dimensional data and categorical features.

The accuracy, precision, recall, and F1 score values indicate that Naive Bayes effectively classifies instances while maintaining a balance between precision and recall.

Gradient Boost:

Gradient Boost is an ensemble learning technique that builds a strong model by sequentially adding weak models in a gradient descent manner.

In our report, Gradient Boost achieved an accuracy of 0.9075, precision of 0.8909, recall of 0.9141, and F1 score of 0.8909. Gradient Boost is robust to overfitting and performs well with complex, nonlinear relationships between features and the target variable.

The performance metrics indicate that Gradient Boost effectively captures the underlying patterns in the data and achieves high accuracy with balanced precision and recall.

CART (Classification and Regression Trees):

CART is a decision tree algorithm that recursively splits the dataset into subsets based on the values of input features.

In our report, CART achieved an accuracy of 0.8238, precision of 0.7812, recall of 0.8107, and F1 score of 0.7812. CART is interpretable, easy to understand, and performs well with both numerical and categorical features.

The performance metrics suggest that CART provides a reasonable balance between accuracy, precision, recall, and interpretability.

Support Vector Machine (SVM):

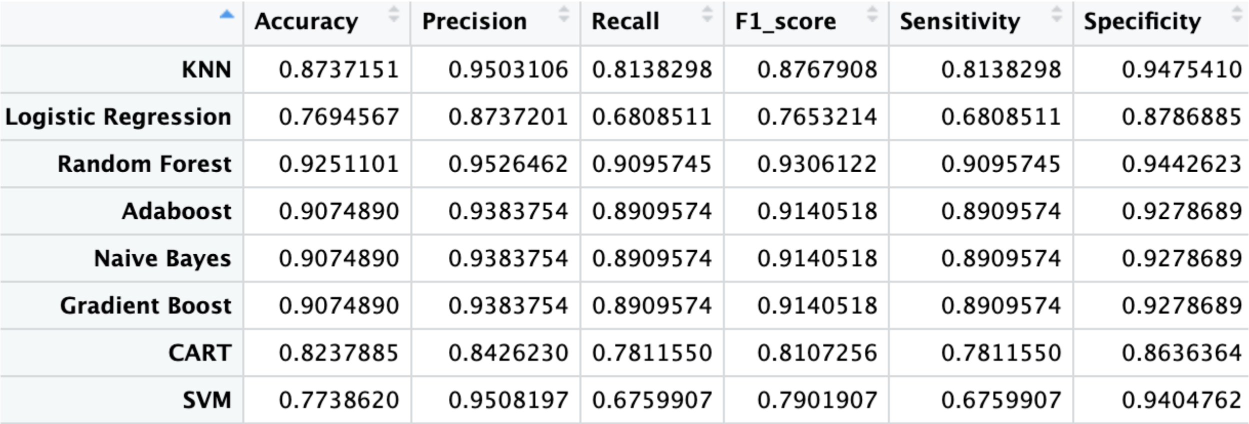
SVM is a supervised learning algorithm that finds the optimal hyperplane to separate classes in a high-dimensional feature space.

In our report, SVM achieved an accuracy of 0.7739, precision of 0.6760, recall of 0.7902, and F1 score of 0.6760. SVM is effective in handling high-dimensional data and performs well in cases where the classes are linearly separable or nearly separable.

The performance metrics indicate that SVM provides reasonable accuracy, but there may be challenges in achieving high precision with this algorithm, especially in cases of class imbalance or overlapping classes.

Validation of Machine learning algorithms

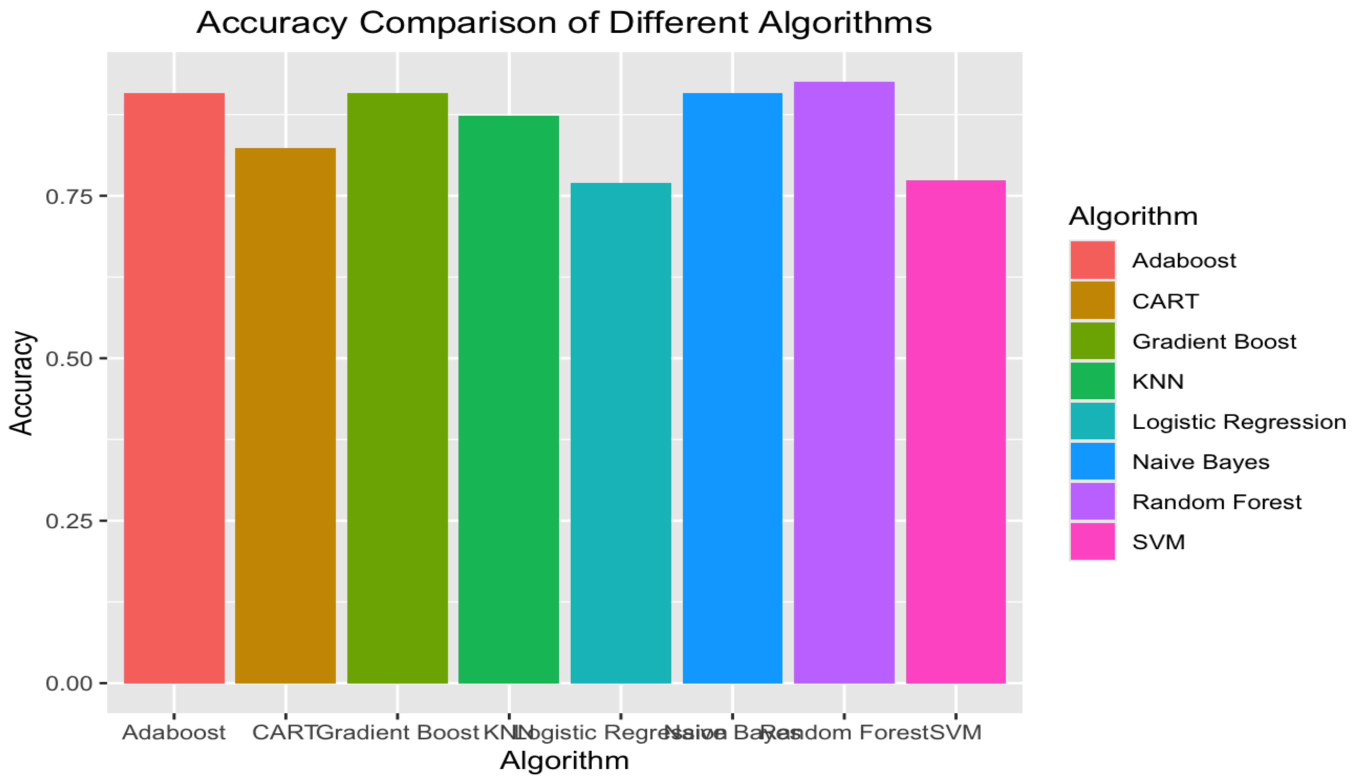
Below table presents a comprehensive statistical metrices applied for validating the eight machine learning algorithms. These metrices include Accuracy, sensitivity, specificity, precision, recall and F1Score. Among the eight machine learning algorithms considered, Random Forest had the highest accuracy, Sensitivity value, precision value, recall and F1 value, K nearest Neighbour had the highest specificity value.



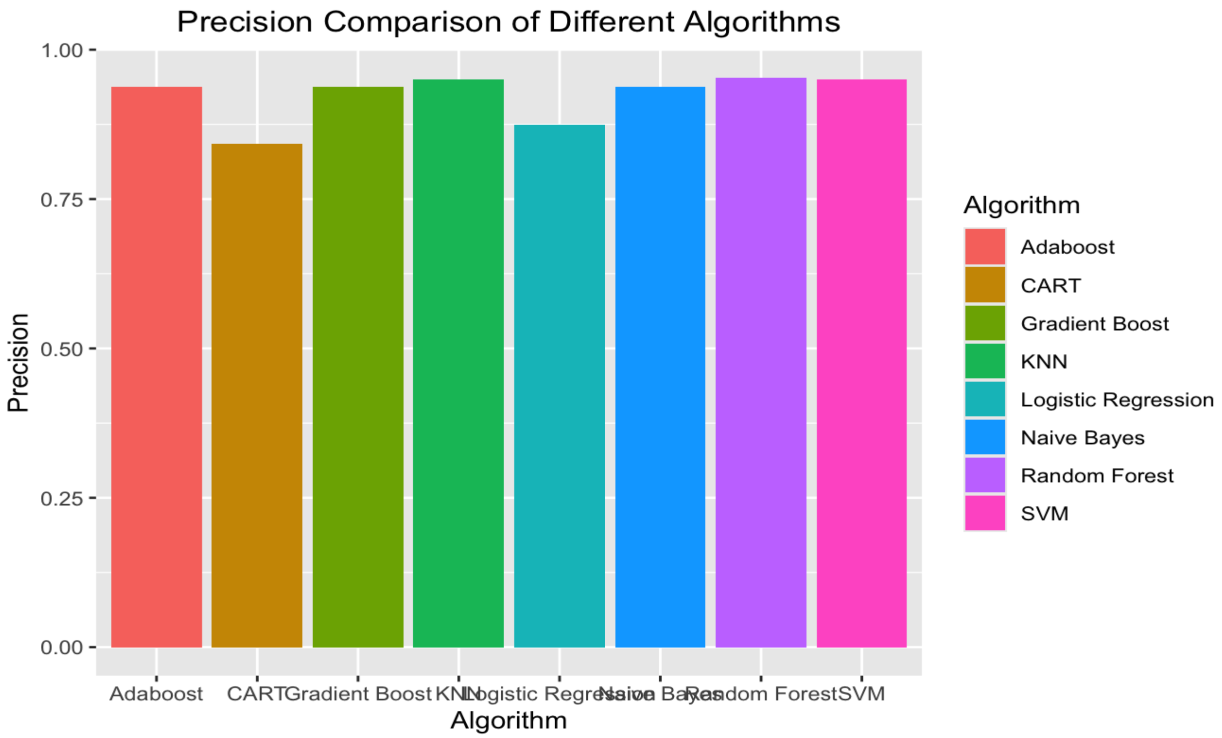
	Accuracy	Precision	Recall	F1_score	Sensitivity	Specificity
KNN	0.8737151	0.9503106	0.8138298	0.8767908	0.8138298	0.9475410
Logistic Regression	0.7694567	0.8737201	0.6808511	0.7653214	0.6808511	0.8786885
Random Forest	0.9251101	0.9526462	0.9095745	0.9306122	0.9095745	0.9442623
Adaboost	0.9074890	0.9383754	0.8909574	0.9140518	0.8909574	0.9278689
Naive Bayes	0.9074890	0.9383754	0.8909574	0.9140518	0.8909574	0.9278689
Gradient Boost	0.9074890	0.9383754	0.8909574	0.9140518	0.8909574	0.9278689
CART	0.8237885	0.8426230	0.7811550	0.8107256	0.7811550	0.8636364
SVM	0.7738620	0.9508197	0.6759907	0.7901907	0.6759907	0.9404762

Comparative Analysis of Varied Machine Learning Algorithms

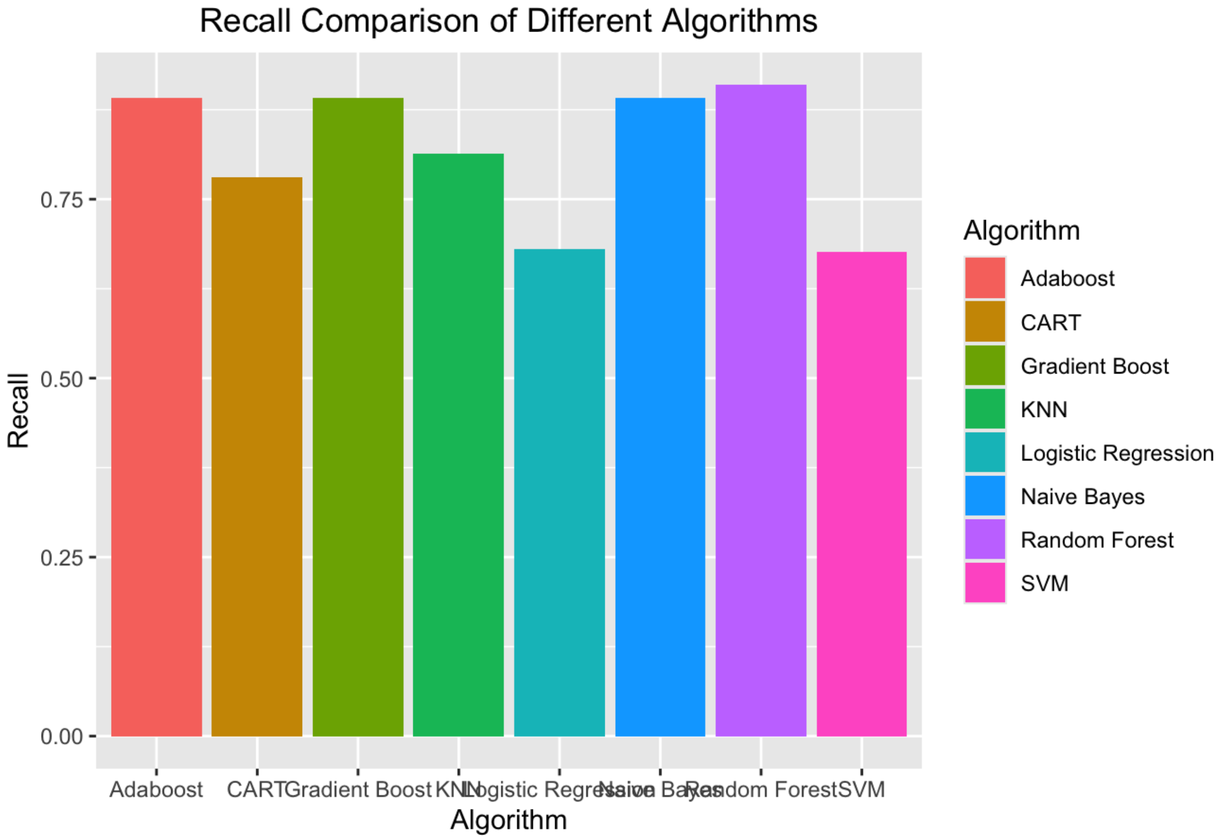
The graphs as presented below to portray the comparative analysis for each metrices as regards the eight machine learning algorithms



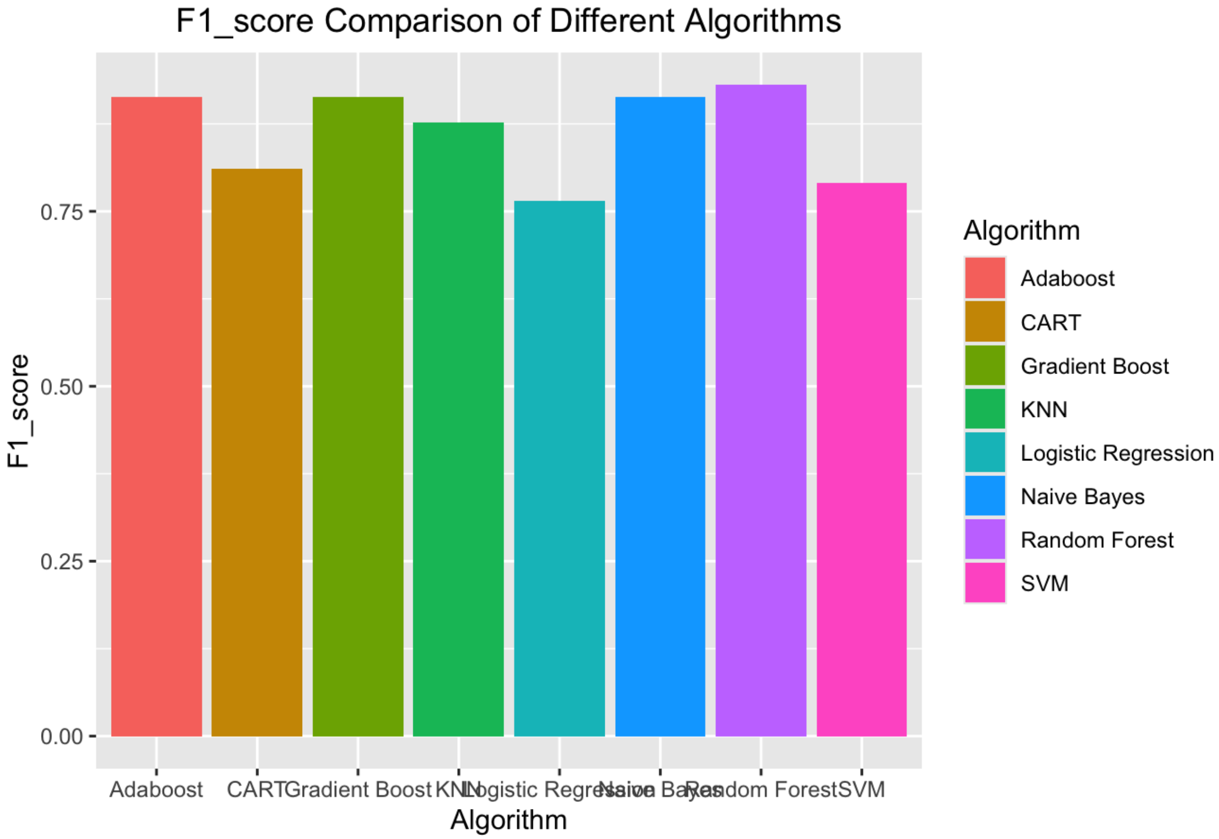
The above Figure depicts the accuracies of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest accuracy value followed by Ada Boost.



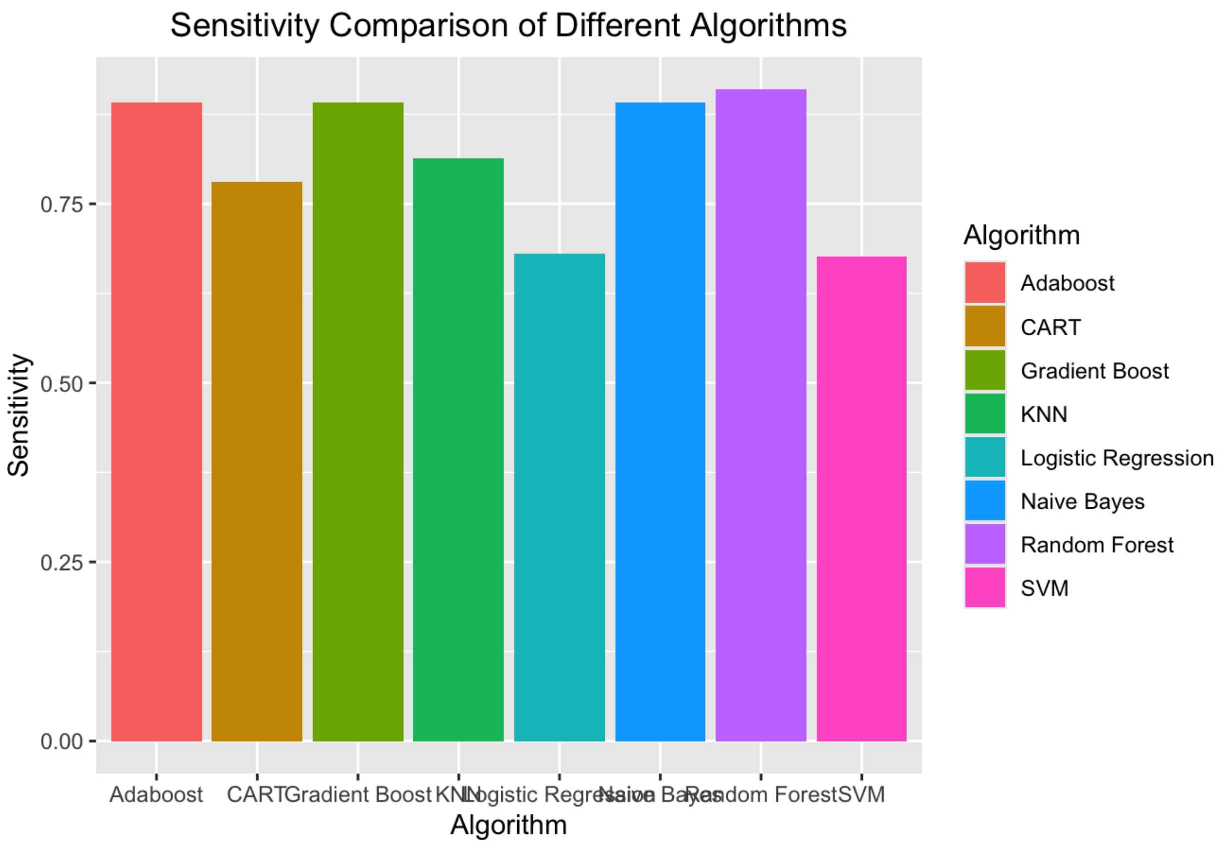
The above Figure depicts the precisions of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest precision value followed by KNN.



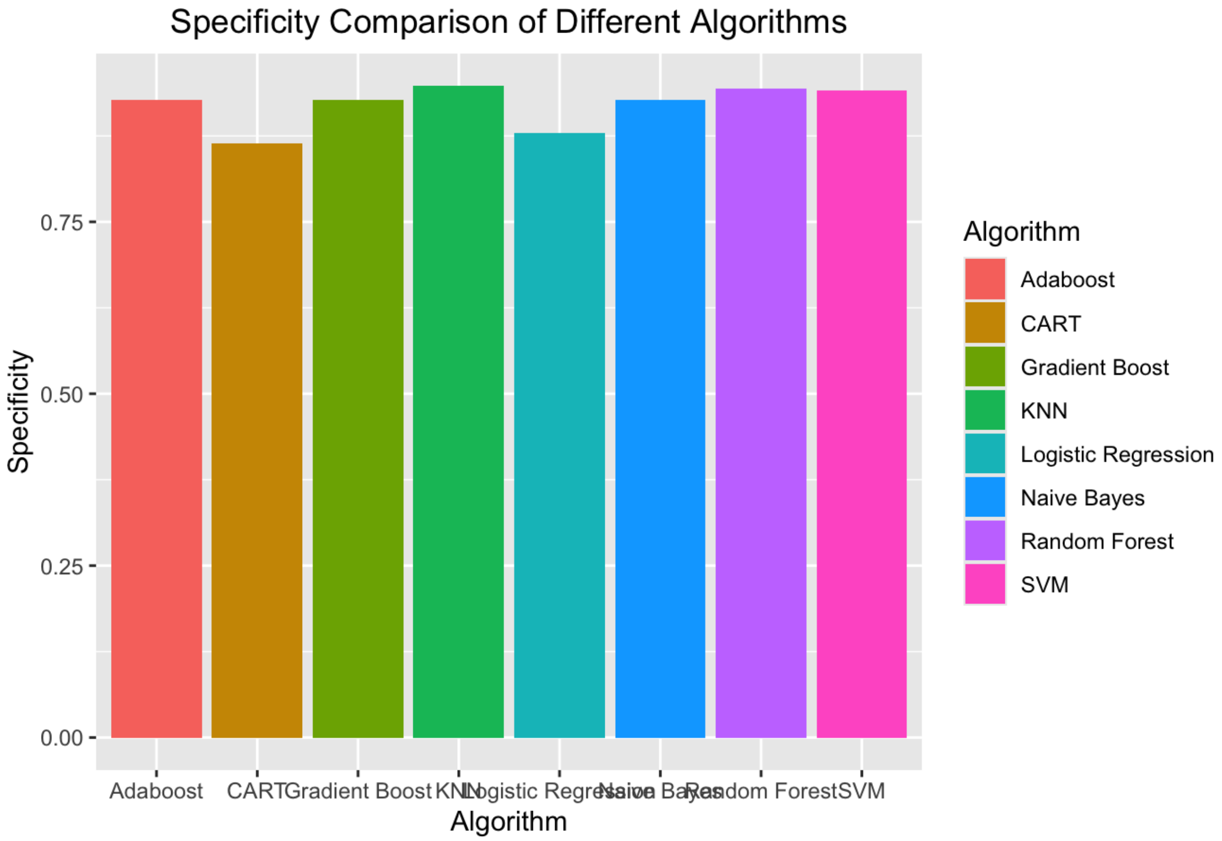
The above Figure depicts the Recall of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest recall value.



The above Figure depicts the F1 scores of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest F1 value.



The above Figure depicts the Sensitivity of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest Sensitivity.



The above Figure depicts the specificity of the eight machine learning algorithms implemented. According to the graph Random Forest has the highest specificity.

Conclusion

This research has successfully implemented eight machine learning algorithms for the classification of Obesity. The implemented machine learning algorithms were categorized into conventional machine learning algorithms and ensemble machine learning algorithms. The presented results and findings have weakened the assertion that ensemble machine learning algorithms outpace conventional machine algorithms, which cannot be ascertained with certainty for this research as pertaining to “boosting”, an ensemble technique which addresses and lessens bias. The research findings have thus paved the way for further studies.