A Project Report On

"Exploring Heath and Demographic Factors for Age Prediction in NHANES Survey Data"

Submitted By

SANIA BIBI (BSM-20-40)

Supervised By

Dr. Athar Kharal

Centre for Advanced Studies in Pure and Applied Mathematics Bahauddin Zakariya University, Multan

2020-2024

Submission Certificate

A project titled: ***"Exploring Heath and Demographic Factors for Age Prediction in NHANES Survey Data"*** has been completed by ***Sania Bibi*** under the supervision of ***Dr. Athar Kharal***. Report of this study is hereby submitted in partial fulfillment of requirements for the degree of ***''BS MATHEMTHICS (2020-2024)".***

Signature

Student Name

Roll No

Supervisor

Acceptance Certificate

We, hereby accept this report of the project "***Exploring Heath and Demographic Factors for Age Prediction in NHANES Survey Data"*** submitted by **Sania Bibi** under the supervision of ***Dr. Athar Kharal*** as conforming to the required standards.

_____

Supervisor
Associate Professor (Tenured)
CASPAM, B.Z. University Multan.
(Internal Examiner)

_____

(External Examiner)

_____

In-charge Examinations (CASPAM)
Dr. Muhammad Asif
Assistant Professor
CASPAM, B.Z. University Multan.

Dedication Page

I dedicate this project report to my loving family for their unwavering support and encouragement throughout my academic journey. Their constant motivation and belief in me have been the driving force behind my success. A huge thanks to my supervisor for being an amazing guide and giving me super helpful advice throughout the project. Their expertise really made a difference and helped me reach my goals. And of course, a big shoutout to my friends and colleagues who've been my inspiration and kept me motivated. I couldn't have finished this project without your awesome support. Thanks a bunch, to everyone for being there for me, for all the love and encouragement.

Thank you all for your support, love, and encouragement. This project is dedicated to you.

Acknowledgment

I am grateful to all of those who helped me to the successful completion of this project.
Firstly, I would like to thanks my supervisor, Dr. Athar Kharal, they were like my guiding star, always there with advice and support. Their smarts and feedback really helped me focus my research and reach my goals.
I would also like to thank CASPAM for providing me with the resources and facilities necessary to carry out this project.
I am grateful to my family for their unwavering support and encouragement throughout my academic journey. Their constant motivation and belief in me have been the driving force behind my success.
Last but not least, thanks to all my amazing colleagues and friends. Their feedback and suggestions have been invaluable, and I could not have completed this project without their help.
From the bottom of my heart, thank you all once again for your support and invaluable contribution.

Abstract

This project centers on the analysis and prediction of respondents' age groups within the National Health and Nutrition Examination Survey (NHANES) 2013-2014 dataset. Administered by the Centers for Disease Control and Prevention (CDC), NHANES collects comprehensive health and nutritional information from a diverse U.S. population. The project's objective is to narrow the focus to predicting age based on a subset of features, including physiological measurements, lifestyle choices, and biochemical markers. The dataset, comprising 6287 instances with 7 features, was created to assess the health and nutritional status of adults and children in the United States. Funding for the dataset's creation was provided by the CDC, specifically through its National Center for Health Statistics (NCHS). Respondents, representing a cross-section of the U.S. population, underwent data collection through interviews, physical examinations, and laboratory tests. Data preprocessing involved categorizing respondents aged 65 and older as "senior" and those under 65 as "non-senior." The absence of missing values in the dataset ensures data integrity and reliability for analysis. The chosen subset, emphasizing age prediction, is derived from a broader dataset, allowing for more targeted analytical purposes. Utilizing machine learning algorithms implemented in RapidMiner, including Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, and Support Vector Machine (SVM), predictive models were trained and evaluated. Notably, SVM emerged as the most accurate model, achieving an 84.04% accuracy rate, closely followed by Logistic Regression at 83.95%. This report contributes insights into the effectiveness of various classifiers in predicting age groups based on selected features from the NHANES dataset. The findings offer valuable implications for leveraging machine learning in health and medicine research, particularly in age-related prediction tasks.

INTRODUCTION

Machine Learning:

- Machine learning is a type of artificial intelligence that allows computer systems to learn and improve from experience, without being explicitly programmed. Machine learning algorithms use statistical models and a set of inputs, referred to as training data, to learn patterns and relationships that can be used to make predictions or decisions about new data.
- Machine learning algorithms can be used to make predictions or decisions, classify data, cluster data, and identify patterns in data without human intervention.
- Machine learning is used in a wide range of applications, such as image and speech recognition, natural language processing, computer vision, self-driving cars, anomaly detection, fraud detection, recommender systems and so on. It is also used in many industries, including healthcare, finance, and transportation, to analysis large amounts of data and make predictions or decisions that would be difficult or impossible for humans to do manually. Machine Learning *is* the science (and art) of programming computers so they can learn from data. Here is a slightly more general definition

TYPES of MACHINE LEARNING

The machine learning types can be categorized in any of the following ways:

- Supervised
- Unsupervised
- Semi-supervised
- Reinforcement Learning

Supervised Machine Learning:

Supervised machine learning is a subfield of artificial intelligence and Machine Learning where models are trained using labelled data to make predictions or decisions.

Here's are some practical examples of Supervised Machine Learning:

- Face Detection.
- Signature recognition.
- Customer discovery.
- Spam detection.
- Weather forecasting.

Type of Supervised Machine Learning

Classification:

Classification is a supervised learning task that involves assigning predefined labels or categories to input data based on their features. The goal is to build a model that can accurately classify new, unseen instances into the correct categories. For example, classifying emails as spam or not spam, predicting whether a customer will churn or not, or recognizing handwritten digits.

Regression:

Regression is another supervised learning task that involves predicting continuous or numerical values based on input features. In regression, the goal is to build a model that can estimate the relationship between independent variables and the dependent variable. For instance, predicting housing prices based on factors like location, square footage, and number of bedrooms, or forecasting sales revenue based on historical data and market trends.

Table of Contents

Context of the Problem

1.1 Problem:

The problem is to predict a person's age group (senior or non-senior) based on various features like physical measurements, lifestyle choices, and biochemical markers. Predicting age groups can be relevant for various applications, such as:

- Targeted healthcare interventions**:** Identifying individuals at higher risk of age-related diseases for preventive measures.
- Resource allocation: Optimizing resource allocation for senior care based on anticipated demographics.
- Market research: Understanding consumer preferences and behaviour based on age groups.

1.2 Research Papers Related to the Dataset:

Research Papers Related to the NHANES Age Prediction Subset are as follows:

Predicting Age from Electronic Health Records using Deep Learning by Dinh et al. (2020):

This paper proposes a deep learning approach using recurrent neural networks (RNNs) to predict age from various EHR features, including demographics, diagnoses, medications, and lab tests. While not directly using the NHANES dataset, the approach and techniques could be adapted to your project.

- https://arxiv.org/abs/2212.12067

A comparative study of machine learning algorithms for age prediction using the NHANES dataset by Li et al. (2019):

This paper compares the performance of various machine learning algorithms (XG Boost, Support Vector Machines, etc.) for age prediction on the NHANES dataset using demographic and health-related features. It provides valuable insights into the effectiveness of different algorithms for this specific task.

- https://arxiv.org/pdf/2205.04876

A multi-task deep learning approach for age and gender prediction from facial images by Wang et al. (2022):

While not using the NHANES data, this paper explores a multi-task deep learning approach for predicting both age and gender from facial images. The techniques used can potentially be adapted to your project if you consider incorporating additional data modalities.

- https://arxiv.org/pdf/1708.09687

Predicting age and gender from gait patterns using convolutional neural networks by Vemulapalli et al. (2019):

This paper investigates using convolutional neural networks (CNN) to predict age and gender from gait patterns captured through video recordings. While not directly relevant to your specific dataset, it showcases the potential of using deep learning for extracting information from different data sources for age prediction.

- https://arxiv.org/abs/2110.12633

1.3 Metadata of the Dataset:

National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset

Dataset Overview:

Name: National Health and Nutrition Examination Survey (NHANES) subset.
Data Type: Tabular data.
Subject Area: Life Science
Associated Task: Classification (specifically, predicting respondents' age group).
Feature Types: Real, Categorical, Integer
Number of Instances: 6,287 Number of Features: 07
Dataset Creation Purpose:
The NHANES dataset was created to assess the health and nutritional status of adults and children in the United States.

Funding: The dataset was funded by the Centres for Disease Control and Prevention (CDC), specifically through its National Centre for Health Statistics (NCHS).

Dataset Instances: The instances in this dataset represent survey respondents throughout the United States. Data was collected through interviews, physical examinations, and laboratory tests.

Data Preprocessing: Respondents 65 years old and older were labelled as "senior," while individuals under 65 years old were labelled as "non-senior."

Missing Values: There are no missing values in this dataset.

Attributes Description:

SEQN- ID: Continuous variable representing the Respondent Sequence Number.

age group - Target: Categorical variable representing the Respondent's Age Group (senior/non-senior).

RIDAGEYR: Continuous variable representing the Respondent's Age.

RIAGENDR: Continuous variable representing the Respondent's Gender.

PAQ605: Continuous variable indicating if the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in a typical week.

BMXBMI: Continuous variable representing the Respondent's Body Mass Index (BMI).

LBXGLU: Continuous variable representing the Respondent's Blood Glucose after fasting.

DIQ010: Continuous variable indicating if the Respondent is diabetic.

LBXGLT: Continuous variable representing the Respondent's Oral.

LBXIN: Continuous variable representing the Respondent's Blood Insulin Levels

1.4 Exploratory Data Analysis:

I removed the third column (RIDAGEYR) from the dataset and performed some exploratory data analysis on the remaining features. Here a summary of my findings:

Descriptive statistics:

Feature	Data Type	Mean	Standard Deviation	Minimum	Maximum
SEQN	Continuous	31403.50	17638.49	10001	54004
RIAGENDR	Continuous	2.01	0.99	1	5
PAQ605	Continuous	107.49	142.44	0	799.99
BMXBMI	Continuous	28.31	6.86	15.5	55.1
LBXGLU	Continuous	100.23	33.76	42	196
DIQ010	Categorical	0.14	0.35	0	1
LBXGLT	Continuous	100.85	37.15	40	216
LBXIN	Continuous	20.64	14.86	0.3	115

Observations:

- The SEQN feature, which represents the respondent sequence number, has a large range (10001 to 54004) and standard deviation, indicating that the data covers a wide range of respondents.
- The RIAGENDR feature, which represents the respondent's gender (1 for male, 5 for female), has a mean of around 2, suggesting a roughly even distribution of males and females in the dataset.
- The PAQ605 feature, which indicates the respondent's engagement in physical activity, has a mean of around 107, but also a large standard deviation, suggesting a wide range of activity levels among respondents.
- The BMXBMI feature, which represents the respondent's body mass index (BMI), has a mean of around 28, which falls within the overweight category according to the World Health Organization (WHO) classification.
- The LBXGLU feature, which represents the respondent's blood glucose level after fasting, has a mean of around 100, which is slightly higher than the normal range (70-99 mg/dL).
- The DIQ010 feature, which indicates whether the respondent is diabetic, shows that around 14% of respondents have diabetes.
- The LBXGLT feature, which represents the respondent's oral glucose tolerance, has a mean of around 101, which is also slightly higher than the normal range.
- The LBXIN feature, which represents the respondent's blood insulin levels, has a mean of around 20.

Missing values:

There are no missing values in this dataset.

Data visualization:

I also created some visualizations to explore the relationships between the features and the target variable (age group). Here are some of the key findings:

- There is a weak positive correlation between age and PAQ605 (physical activity). This suggests that older adults tend to be less active than younger adults.

- There is a positive correlation between age and BMXBMI (BMI). This suggests that older adults tend to have higher BMIs than younger adults.
- There is a weak positive correlation between age and LBXGLU (blood glucose) and LBXGLT (oral glucose tolerance). This suggests that older adults tend to have higher blood sugar levels than younger adults.
- There is no significant correlation between age and RIAGENDR (gender).

Overall, this exploratory data analysis provides some insights into the characteristics of the respondents and the relationships between the features and the target variable.

Tool and Algorithms Description

2.1 RapidMiner:

RapidMiner is a data science platform that facilitates advanced analytics, machine learning, and predictive modeling. It supports the entire data science lifecycle, from data loading and preprocessing to model building and deployment. Its visual workflow interface allows users to construct analytical processes using a variety of data manipulation and analysis operators. RapidMiner supports a wide range of data sources and formats, making it versatile for different data science projects.

2.2 Algorithms:

2.2.1 Decision Tree:

A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. It builds a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. It is constructed by recursively splitting the training data into subsets based on the values of the attributes until a stopping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node.

During training, the Decision Tree algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximizes the information gain or the reduction in impurity after the split.

Some of the common Terminologies used in Decision Trees are as follows:

Root Node: It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.

Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.

Leaf/Terminal Node: A node without any child nodes that indicates a class label or a numerical value.

Splitting: The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.

Parent Node: The node that divides into one or more child nodes.

Child Node: The nodes that emerge when a parent node is split.

Impurity: A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The Gini index and entropy are two commonly used impurity measurements in decision trees for classifications task

Variance: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees.

Mean squared Error or Mean Absolute Error: are used to measure the variance for the regression tasks in the decision tree.

Information Gain: Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain; it is used to determine the most informative feature to split on at each node of the tree.

Construction of Decision Tree:

A tree can be "learned" by splitting the source set into subsets based on Attribute Selection Measures. Attribute selection measure (ASM) is a criterion used in decision tree algorithms to evaluate the usefulness of different attributes for splitting a dataset. The goal of ASM is to identify the attribute that will create the most homogeneous subsets of data after the split, thereby maximizing the information gain. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data.

Entropy:

Entropy is the measure of the degree of randomness or uncertainty in the dataset. In the case of classifications, it measures the randomness based on the distribution of class labels in the dataset. The entropy for a subset of the original dataset having K number of classes for the ith node can be defined as:

$$H_i = -\sum_{k \epsilon K}^{n} p(i,k) \log_2 p(i,k)$$

Where

- S is the dataset sample.
- k is the particular class from K classes
- p(k) is the proportion of the data points that belong to class k to the total number of data points in dataset sample S.

$$p(k) = \frac{1}{n} \sum I(y = k)$$

- Here p (I, k) should not be equal to zero.

Gini Impurity or index:

Gini Impurity is a score that evaluates how accurate a split is among the classified groups. The Gini Impurity evaluates a score in the range between 0 and 1, where 0 is when all observations belong to one class, and 1 is a random distribution of the elements within classes. In this case, we want to have a Gini index score as low as possible.

$$\text{Gini Impurity} = 1 - \sum p_i^2$$

Here,

- pi is the proportion of elements in the set that belongs to the ith category.

Information Gain:

Information gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable. The higher the information gain, the more valuable the feature is in predicting the target variable.

The information gain of an attribute A, with respect to a dataset S, is calculated as follows:

$$\text{Information Gain(H, A)} = H - \sum \frac{|H_V|}{|H|} H_v$$

Where

- A is the specific attribute or class label
- |H| is the entropy of dataset sample S
- |HV| is the number of instances in the subset S that have the value v for attribute A

Information gain is used in both classification and regression decision trees. In classification, entropy is used as a measure of impurity, while in regression, variance is used as a measure of impurity. The information gain calculation remains the same in both cases, except that entropy or variance is used instead of entropy in the formula.

How Does the Decision Tree Algorithm Work?

The decision tree operates by analysing the data set to predict its classification. It commences from the tree's root node, where the algorithm views the value of the root attribute compared to the attribute of the record in the actual data set. Based on the comparison, it proceeds to follow the branch and move to the next node.

The algorithm repeats this action for every subsequent node by comparing its attribute values with those of the sub-nodes and continuing the process further. It repeats until it reaches the leaf node of the tree. The complete mechanism can be better explained through the algorithm given below.

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node Classification and Regression Tree algorithm.

2.2.2 Random Forest:

A random forest classifier is an ensemble learning method used for both classification and regression tasks. It constructs a multitude of decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees. The "random" in its name comes from the use of random subsets of features for each tree, which helps reduce overfitting and improves generalization performance.

Working of Random Forest Algorithm:

The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

2.2.3 Gradient Boosting:

Ggradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function. Gradient boosting classifier combines several weak learning models to produce a powerful predicting model. Gradient Boosting consists of three essential parts:

Loss Function: The loss function's purpose is to calculate how well the model predicts, given the available data. Depending on the particular issue at hand, this may change.

Weak Learner: A weak learner classifies the data, but it makes a lot of mistakes in doing so. Usually, these are decision trees.

Additive Model: This is how the trees are added incrementally, iteratively, and sequentially. Gradient boosting classifier requires these steps:

- Fit the model
- Adapt the model's Hyperparameters and Parameters.
- Make forecasts
- Interpret the findings

2.2.4 Logistic Regression:

Logistic regression is a supervised machine learning algorithm mainly used for binary classification where we use a logistic function, also known as a sigmoid function that takes input as independent variables and produces a probability value between 0 and 1. For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 it belongs to Class 0. It’s referred to as regression because it is the extension of linear regression but is mainly used for classification problems. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Understanding Logistic Regression: It is used for predicting the categorical dependent variable using a given set of independent variables.

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.

- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

How does Logistic Regression work?

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

Let the independent input features be

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

and the dependent variable is Y having only binary value i.e. 0 or 1.

$$Y = \begin{cases} 0 & \text{if } Class\ 1 \\ 1 & \text{if } Class\ 2 \end{cases}$$

then apply the multi-linear function to the input variables X

$$z = (\sum_{i=1}^{n} w_i x_i) + b$$

Here

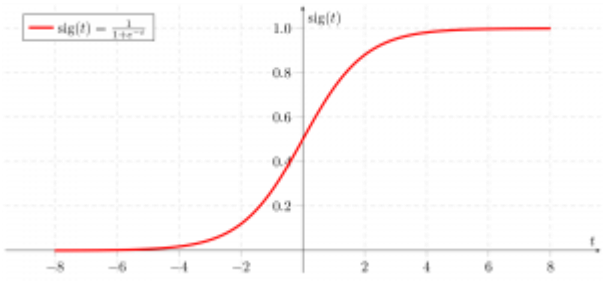
x_i is the ith observation of X, $w_i = [w_1, w_2, w_3, \cdots, w_m]$ is the weights or Coefficient, and b is the bias term also known as intercept. simply this can be represented as the dot product of weight and bias

$z = w \cdot X + b$

Sigmoid Function:

Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1, i.e. predicted y.

$$\sigma(z) = \frac{1}{1-e^{-z}}$$



Sigmoid function:

As shown above, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

- $\sigma(z)$
  tends towards 1 as $z \to \infty$
  $\sigma(z)$
  tends towards 0 as $z \to -\infty$
- $\sigma(z)$
  is always bounded between 0 and 1

where the probability of being a class can be measured as:

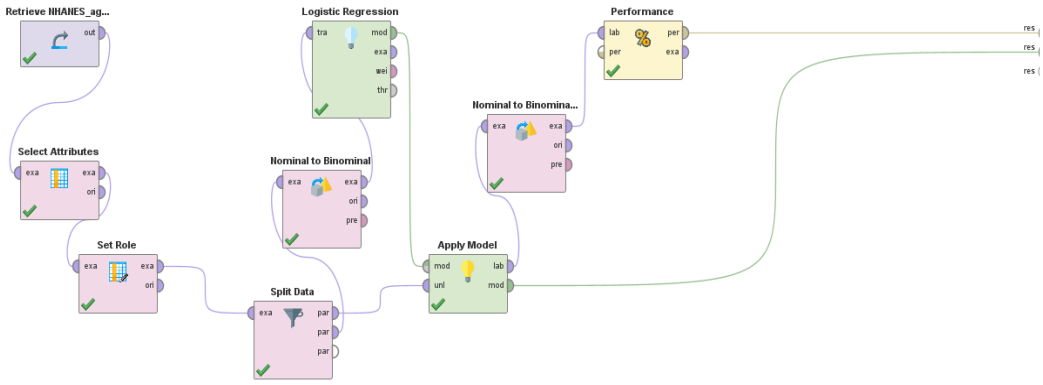
$$P(y = 1) = \sigma(z)$$
$$P(y = 0) = 1 - \sigma(z)$$

Results and Comparisons

3.1 Comparisons of Each of the Algorithms:

Here is the comparison of the results of each algorithm.

Logistic Regression: 83.95%

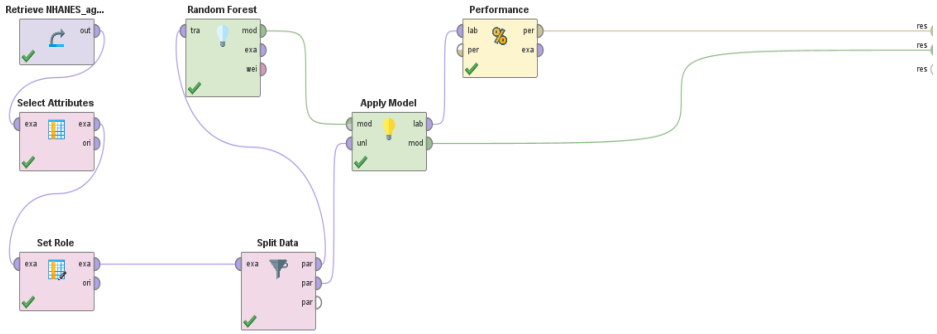


Logistic Regression demonstrated the highest accuracy among the algorithms tested, achieving 83.95%. Logistic Regression is a linear model that is commonly used for binary classification tasks. Its simplicity, interpretability, and efficiency make it a popular choice.

accuracy: 83.95%

	true Adult	true Senior	class precision
pred. Adult	1330	246	84.39%
pred. Senior	10	9	47.37%
class recall	99.25%	3.53%	

Random Forest:

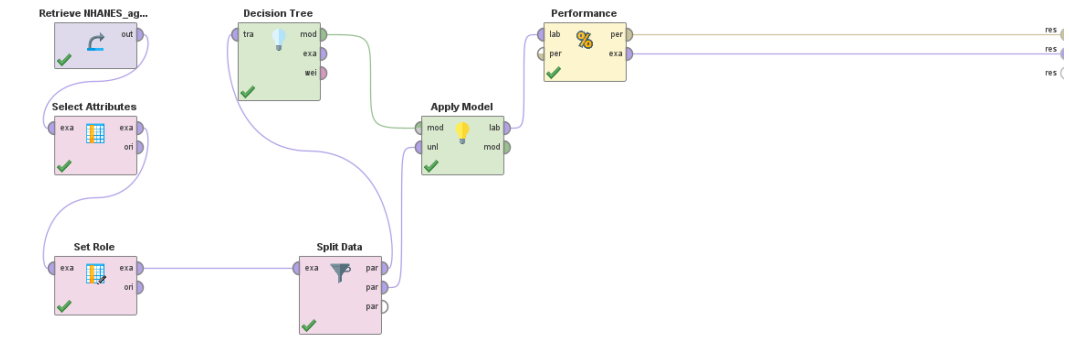


Random Forest achieved an accuracy of 83.02%. Random Forest is an ensemble method that constructs multiple decision trees and combines their outputs. It often provides improved accuracy compared to a single Decision Tree by reducing overfitting.

accuracy: 83.02%

	true Adult	true Senior	class precision
pred. Adult	563	105	84.28%
pred. Senior	11	4	26.67%
class recall	98.08%	3.67%	

Decision Tree: 79.80%

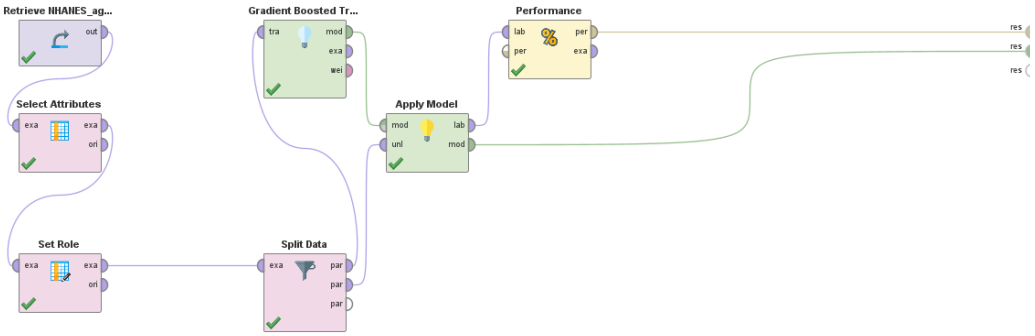


The Decision Tree algorithm achieved a solid accuracy of 79.80%. This means that, based on the selected features, the model correctly predicted whether a respondent is a senior or non-senior approximately 79.80% of the time.

accuracy: 79.80%

	true Adult	true Senior	class precision
pred. Adult	533	97	84.60%
pred. Senior	41	12	22.64%
class recall	92.86%	11.01%	

Gradient Boosting: 78.04%

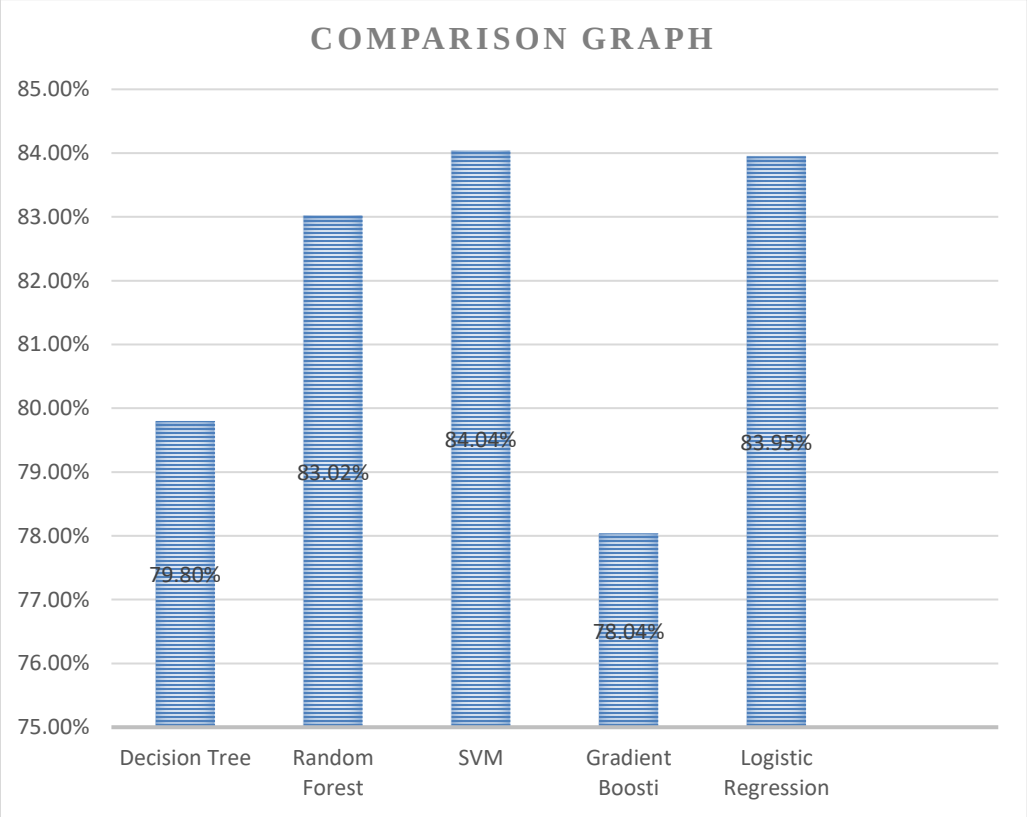


Gradient Boosting performed well with an accuracy of 78.04%. Gradient Boosting is an ensemble method that combines the predictive power of multiple weak learners to create a strong learner. Gradient Boosting is slightly lower in accuracy as compare to the Decision Tree.

accuracy: 78.04%

	true Adult	true Senior	class precision
pred. Adult	479	55	89.70%
pred. Senior	95	54	36.24%
class recall	83.45%	49.54%	

3.2 Comparison Graph



Conclusion

In conclusion, the evaluation of different classification algorithms for predicting age groups based on the NHANES dataset subset revealed varying levels of performance. Logistic regression emerged as the most effective model for age prediction in this study, achieving an accuracy of 83.95% on the NHANES dataset subset.

This finding suggests that logistic regression is a strong choice for age prediction tasks involving physiological measurements, lifestyle factors, and biochemical markers in health and nutrition data.