

**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**A REPORT**

**On**

**"A Study On Regression Based Prediction Model"**

*Submitted in partial fulfillment of the requirements for the Course Introduction to Machine Learning (4CSPL2041) in*

**Bachelor of Technology In**

**Computer Science and Engineering**

*SoET, CMR University, Bangalore*

Submitted by:

**Indhu S (21BBTCS097)**

**Madhushree K M (21BBTCS122) Manasa B M (21BBTCS124)**

**Under the Supervision**:

**Dr. Manjunath C R**

Professor

**Department of Computer Science and Engineering**

**Off Hennur - Bagalur Main Road,**

**Near Kempegowda International Airport, Chagalahatti, Bangalore, Karnataka-562149**

**2023-2024**



**SCHOOL OF ENGINEERING AND TECHNOLOGY**

Chagalahatti, Bengaluru, Karnataka- 562149

Department of Computer science and engineering

CERTIFICATE

This is to certify that the study Report entitled **“ A Study On Regression Based Prediction Model ”**, is a record of work successfully carried out by **Indhu S (21BBTCS097), Madhushree K M (21BBTCS122), Manasa B M (21BBTCS124)** in partial fulfilment of the requirement for the course **INTRODUCTION TO MACHINE LEARNING (4CSPL2041)** of Bachelor of Technology in Computer Science and Engineering, SOET, CMR University, Bangalore during the academic year 2023- 24, under the supervision and guidance of **Dr. MANJUNATH C.R,** Professor, CSE, SOET, CMR University.

Signature

**Dr. MANJUNATH C.R,**

**Professor, Dept of CSE, SOET, CMR University.**



|  |  |  |
| --- | --- | --- |
| **Chapter**  **No** | **Title** | **Page No** |
|  | **ABSTRACT** | **1** |
| **1** | **INTRODUCTION** | **2-6** |
|  | **1.1 Background and Context** | **2-6** |
|  | **1.2 Objectives of the study** | **6** |
| **2** | **MACHINE LEARNING MODEL** | **7-20** |
|  | **2.1 Linear Regression** | **7-9** |
|  | **2.2 Logistic Regression** | **9-10** |
|  | **2.3 Random Forest** | **10-12** |
|  | **2.4 XG Boost** | **12-14** |
|  | **2.5 Support Vector Machine (SVM)** | **14-15** |
|  | **2.6 MLP** | **15-16** |
|  | **2.7 K-Nearest Neighbor (KNN)** | **16-17** |
|  | **2.8 Naïve Bayes** | **17-19** |
| **3** | **INTERPRETATION OF MODELS** | **20** |
| **4** | **CONCLUSION** | **21** |
| **5** | **REFERENCES** | **22-23** |



|  |  |  |
| --- | --- | --- |
| **Figure no** | **Title** | **Page no** |
| 1.1 | Body Fat Estimation Uing Machine Learning | 3 |
| 2.1.1 | Linear Regression Model Of Body Fat Prediction | 8 |
| 2.2.1 | Logistic Regression Model | 9 |
| 2.3.1 | Random Forest Prediction Model | 10 |
| 2.4.1 | XG-Boost Prediction Model | 13 |
| 2.5.1 | SVM Model | 14 |
| 2.6.1 | MLP Model | 16 |
| 2.8.1 | Naive Bayes Model | 18 |



|  |  |  |
| --- | --- | --- |
| **Table no** | **Title** | **Page no** |
| 3.1 | Analyzing the Models | 20 |

## ABSTRACT

Body fat stands as a critical public health issue, posing significant risks for various serious diseases. While several methods exist for estimating body fat to gauge obesity, these often entail expensive tests and specialized equipment. Therefore, accurate prediction of body fat percentage through easily accessible body measurements becomes paramount for effectively assessing obesity and its associated health implications. This study delves into the realm of feature extraction, examining its efficacy in predicting body fat. By scrutinizing the distinctive characteristics of different features, such as various body measurements, the research endeavors to discern the impact of feature extraction on prediction accuracy. A rigorous evaluation process scrutinizes three distinct feature extraction methodologies alongside four widely recognized prediction models. Drawing upon data from two real-world body fat datasets, the study meticulously assesses the performance of these models, focusing on key metrics like mean absolute error, standard deviation, root mean square error, and robustness. The findings unequivocally demonstrate that integrating feature extraction significantly enhances the predictive capability of these models for body fat estimation. Not only do the results confirm the efficacy of feature extraction as a preprocessing step in body fat prediction, but they also underscore its pivotal role in refining the accuracy and robustness of predictive models in this domain. Moreover, statistical analyses corroborate the substantial improvement afforded by feature extraction, shedding light on its tangible benefits in bolstering prediction performance. Notably, the study observes that augmenting the number of extracted features yields marginal yet discernible enhancements in the predictive accuracy of models, further accentuating the importance of comprehensive feature selection.

In essence, this study contributes valuable insights into the efficacy of feature extraction techniques in body fat prediction and establishes a robust baseline for future research endeavors in related fields. By illuminating the significance of feature extraction as a catalyst for refining prediction models, the findings pave the way for more nuanced approaches to obesity assessment and management, ultimately fostering advancements in public health interventions to combat the obesity epidemic.

**Keywords:** Machine learning, body fat percentage prediction, linear regression, Random Forest(RF), Decision Tree, XG - Boost, MLP, healthcare, fitness, wellness, obesity, personalized assessment, intervention strategies.

## CHAPTER – 1

### Background and Context

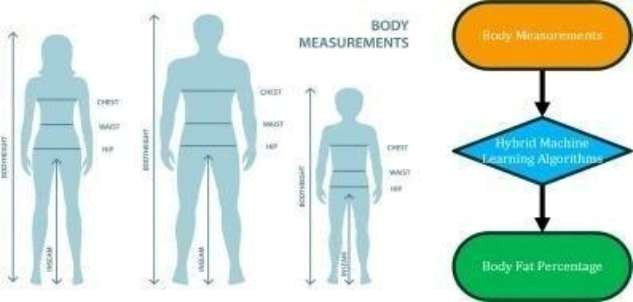
## INTRODUCTION

Body fat prediction using machine learning (ML) models encompasses a diverse array of techniques tailored to accurately estimate body fat percentage and distribution. Linear regression serves as a foundational model, employing a straightforward approach to modeling the relationship between input features and body fat percentage. Random forest and decision trees offer more complex solutions, capable of capturing nonlinear relationships and interactions among features, making them suitable for analyzing intricate datasets. Anomaly detection techniques play a crucial role in identifying outliers or irregular patterns in body fat data, aiding in data preprocessing and outlier removal. XG Boost, a gradient boosting algorithm, further enhances predictive accuracy by iteratively improving upon weak learners. Multilayer perceptron (MLP) neural networks leverage interconnected layers of neurons to learn complex patterns in body fat data, while support vector machines (SVM) construct hyperplanes to separate data points into different classes based on feature vectors. Lastly, artificial neural networks (ANN) offer a versatile framework for body fat prediction, capable of learning intricate mappings between input features and body fat percentage through layers of interconnected neurons, often yielding state-of-the-art performance in predictive tasks. These ML models collectively contribute to advancing our understanding of body composition and informing personalized health and fitness interventions.

Body fat prediction using machine learning (ML) models not only involves accurately estimating body fat percentage and distribution but also addresses various challenges and considerations inherent to the field of health and fitness sciences. Beyond the technical aspects, there are critical factors such as data quality, feature selection, model interpretability, and ethical implications that impact the efficacy and responsible deployment of these models. Ensuring the quality and representativeness of the training data is essential to develop robust ML models capable of generalizing well to unseen data and diverse populations. Feature selection techniques help identify the most informative variables for body fat prediction, ensuring that the model focuses on relevant factors while mitigating the risk of overfitting. Moreover, interpretable models such as decision trees and linear regression not only provide accurate predictions but also offer insights into the factors influencing body fat levels, facilitating a deeper understanding of the underlying mechanisms. Ethical considerations, including data privacy, algorithm bias, and the potential impact on individuals' well-being, must be carefully addressed to ensure the responsible and equitable deployment of ML-based body fat prediction models in real-world settings. By integrating technical expertise with ethical awareness and domain knowledge, these ML models can contribute to advancing our understanding of body composition, promoting healthier lifestyles, and reducing the burden of obesity-related diseases on a global scale.

Machine learning (ML) represents a transformative paradigm in the realm of artificial intelligence (AI) and data science, enabling computers to learn from data and improve their performance over time without explicit programming. At its core, machine learning aims to develop algorithms and models that can automatically identify patterns, make predictions, and derive insights from complex datasets. This field encompasses a diverse array of techniques and methodologies, ranging from traditional statistical methods to cutting-edge deep learning architectures. The driving force behind the rapid growth of machine learning is the exponential increase in data availability and computational power, coupled with advancements in algorithm development and optimization techniques.

ML algorithms can broadly categorized into supervised learning, unsupervised learning, semi- supervised learning, and reinforcement learning, each tailored to different types of learning tasks and data modalities. Machine learning has emerged as a powerful tool in predicting body fat in humans, revolutionizing how we understand and manage body composition. By leveraging large datasets and sophisticated algorithms, ML enables precise and personalized predictions of body fat percentage, facilitating tailored interventions and improving overall health outcomes. The integration of ML techniques into body fat prediction models has significantly enhanced their accuracy, reliability, and applicability across diverse populations and settings. At the core of ML-based body fat prediction lies the utilization of predictive models trained on extensive datasets comprising anthropometric measurements, demographic information, lifestyle factors, and biomarkers. These models employ various algorithms, ranging from traditional regression methods to advanced machine learning techniques such as artificial neural networks (ANNs), support vector machines (SVMs), random forests, and gradient boosting machines (GBMs). Through iterative learning processes, these algorithms identify complex patterns and relationships within the data, enabling them to make accurate predictions of body fat percentage based on input features.



**Figure 1.1 : Body Fat Estimation Using Machine learning**

One of the key advantages of ML in body fat prediction is its ability to capture non-linear relationships and interactions among predictor variables, which may not be easily discernible using traditional statistical approaches. ML models can uncover intricate associations between anthropometric measurements, demographic factors, and lifestyle habits, allowing for a more comprehensive understanding of the factors influencing body fat composition. Moreover, ML facilitates the integration of diverse data sources, including wearable devices, electronic health records, and genetic information, enriching the predictive capabilities of body fat prediction models. ML-based body fat prediction models offer several practical applications across healthcare, fitness, and wellness domains. In clinical settings, these models aid healthcare providers in assessing body composition, monitoring changes over time, and identifying individuals at risk of obesity-related complications.

By accurately predicting body fat percentage, ML enables clinicians to tailor interventions, such as personalized diet and exercise plans, to meet the specific needs of patients. Additionally, ML-driven

body fat prediction models empower individuals to take proactive steps towards improving their health and well-being by providing actionable insights and personalized recommendations based on their unique characteristics and lifestyle factors.

Furthermore, ML-based body fat prediction has significant implications for population health management and public health policy. By analyzing large-scale datasets from epidemiological studies and national health surveys, ML models can identify trends, disparities, and risk factors associated with obesity and related comorbidities. This information can inform targeted interventions, preventive strategies, and policy initiatives aimed at reducing the prevalence of obesity and improving population health outcomes. Additionally, ML enables the development of predictive models that can forecast future trends in body fat composition, helping policymakers allocate resources and plan interventions effectively. In conclusion, machine learning plays a pivotal role in predicting body fat in humans, offering advanced analytical tools and predictive models that enhance our understanding of body composition and inform personalized interventions. By harnessing the power of ML algorithms and large-scale data sources, researchers, clinicians, and policymakers can develop accurate, reliable, and actionable predictions of body fat percentage, ultimately contributing to improved health outcomes and enhanced quality of life for individuals and populations alike.

Obesity, a burgeoning global health crisis, transcends mere physical appearance to become a profound medical challenge of our times. Characterized by an accumulation of excess body fat, obesity engenders a cascade of health complications, including but not limited to cardiovascular diseases, diabetes, musculoskeletal disorders, depression, and certain cancers. Paradoxically, extremely low body fat levels harbor their own perils, posing risks such as increased vulnerability to infections, pubertal delays, osteoporosis, and heightened surgical complications. The intricate interplay between body fat levels and health underscores the critical need for precise predictive models that can delineate between excess and insufficient fat levels, thereby guiding preventative measures and therapeutic interventions effectively. Despite the wealth of medical data inundating contemporary healthcare systems, the analysis thereof poses formidable challenges. The curse of dimensionality looms large, exacerbated by the omnipresent specter of redundant, irrelevant, or outright noisy features within datasets sourced from an array of sensors, electronic medical health records, smartphone applications, and insurance records. In this labyrinth of data, the path to discerning meaningful insights appears fraught with complexity.

Enter feature extraction—a beacon of hope amidst the data deluge. Feature extraction methods such as Factor Analysis (FA), Principal Component Analysis (PCA), and Independent Component Analysis (ICA) emerge as stalwart allies, offering respite from the quagmire of dimensionality. Through the alchemy of feature extraction, datasets undergo metamorphosis, shedding extraneous features while preserving the nuggets of essential information that illuminate the path towards predictive accuracy. The fusion of feature extraction techniques with the formidable prowess of machine learning (ML) algorithms heralds a new era in predictive modeling for obesity. Studies abound showcasing the symbiotic relationship between feature extraction and ML, with methodologies ranging from Artificial Neural Networks (ANNs) to Support Vector Machines (SVMs), Random Forests (RF), andextreme Gradient Boosting (XGBoost). These sophisticated algorithms, fueled by curated datasets enriched through feature extraction, hold the promise of unlocking unprecedented insights into the labyrinthine landscape of obesity prediction.

As we embark on this journey of exploration and discovery, guided by the twin beacons of feature extraction and machine learning, we envisage a future where the specter of obesity is met with a formidable arsenal of predictive tools. Through rigorous validation and relentless pursuit of

innovation, we aspire to sculpt a future where obesity's ominous shadow is banished, and the beacon of health shines bright for generations to come. Body fat in men plays a crucial role beyond mere aesthetics, significantly influencing overall health and well-being. Unlike women, who tend to distribute fat more evenly throughout their bodies, men commonly accumulate fat in specific areas, particularly around the abdomen. This abdominal fat, known as visceral fat, is metabolically active and poses a greater risk to health compared to fat stored in other areas. The distribution of body fat in men, often characterized by an "apple-shaped" physique, has significant implications for health outcomes. Excess abdominal fat has been strongly linked to an increased risk of various chronic conditions, including cardiovascular diseases such as heart disease and stroke.

This correlation is attributed to the close proximity of visceral fat to vital organs, which can disrupt their function and contribute to systemic inflammation and insulin resistance. Moreover, excessive body fat in men is associated with hormonal imbalances, particularly reduced levels of testosterone. This hormonal disruption can have far-reaching effects on men's health, impacting fertility, libido, muscle mass, and overall vitality. Additionally, abdominal obesity is a key component of metabolic syndrome, a cluster of conditions including high blood pressure, elevated blood sugar levels, and abnormal cholesterol levels, which collectively elevate the risk of cardiovascular disease and type 2 diabetes. Addressing body fat in men requires a comprehensive approach that encompasses dietary modifications, regular physical activity, and lifestyle changes. Adopting a balanced diet rich in nutrient-dense foods, coupled with portion control and moderation of calorie intake, is crucial for managing body weight and reducing abdominal fat accumulation.

Engaging in regular exercise, including both aerobic and resistance training, not only promotes fat loss but also helps maintain muscle mass and metabolic health. Furthermore, adopting healthy lifestyle habits such as adequate sleep, stress management, and avoiding excessive alcohol consumption can further support efforts to achieve and maintain a healthy body composition. In conclusion, body fat in men extends far beyond cosmetic concerns, exerting a profound impact on health and longevity. By understanding the distribution patterns, health risks, and management strategies associated with body fat, men can take proactive steps to optimize their health and reduce the risk of chronic diseases associated with excess adipose.

### Problem Statement

Develop a machine learning model to predict body fat percentage based on various anthropometric measurements and demographic information. The model should accurately estimate body fat percentage using features such as age, gender, weight, height, waist circumference, hip circumference, and possibly other relevant variables. This predictive tool aims to assist individuals, healthcare providers, and fitness professionals in assessing body composition more accurately and efficiently, facilitating personalized health and fitness recommendations. Developing accurate machine learning models to predict body fat percentage is the project's goal.

Various techniques suchas linear regression, random forest, and neural networks will be explored. Challenges include ensuring data quality, selecting relevant features, and addressing ethical considerations. Success will enhance our understanding of body composition and inform personalized

health interventions. The project's focus is on improving predictive accuracy while considering practical and ethical implications.

### Methodology

1. **Data Collection**: Gather a dataset with anthropometric measurements and demographic info,including body fat percentage.
2. **Data Preprocessing**: Clean the data, handle missing values and outliers, and engineer relevant features like BMI.
3. **Splitting the Data**: Divide the data-set into training, validation, and test sets.
4. **Model Selection**: Choose regression algorithms like linear regression or decision trees.
5. **Model Training**: Train the selected model on the training data.
6. **Model Evaluation**: Assess model performance using validation set, comparing predicted values with actual.
7. **Model Fine-tuning**: Adjust hyper parameters to improve performance.
8. **Final Model Evaluation**: Evaluate the final model using the test set for unbiased performanceestimation.
9. **Deployment**: Deploy the model for real-world use.
10. **Monitoring and Maintenance**: Continuously monitor and update the model as needed.

### Objectives of the study

* Evaluate the effectiveness of feature extraction techniques in predicting body fat percentage.
* Utilize a dataset comprising body fat estimates and various body circumference measurements for 252 men.
* By using a linear regression approach and design and analysis of the impact on prediction model performance.

## CHAPTER - 2

# PREDICTION MODELS

### Linear Regression Model

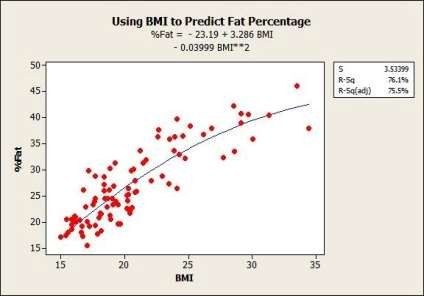
**Regression analysis** is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variables (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.Regression prediction models are foundational tools in statistical analysis, offering insights into the relationships between variables and facilitating predictions of continuous outcomes. At the heart of these models lies the regression equation, which mathematically encapsulates the association between predictor variables and the outcome variable of interest. This equation serves as the blueprint for understanding how changes in the predictor variables correspond to changes in the outcome variable. Through the estimation of regression coefficients, the model quantifies the strength and direction of these relationships, providing valuable insights into the underlying dynamics of the data.

Regression prediction models are versatile, finding applications across a wide range of domains, including finance, economics, healthcare, and engineering. In finance, regression models are used to predict stock prices, analyze market trends, and assess investment strategies. In healthcare, they aid in forecasting patient outcomes, identifying risk factors for diseases, and optimizing treatment plans. Similarly, in engineering, regression models are instrumental in predicting the performance of systems, optimizing processes, and guiding decision-making in areas such as manufacturing and transportation. The success of regression prediction models hinges on several key factors, including the selection of appropriate predictor variables, the quality of the data, and adherence to underlying assumptions. Furthermore, model evaluation techniques, such as assessing goodness-of-fit and validating predictive performance, are critical for ensuring the reliability and robustness of the model.

In recent years, advancements in computational techniques and the availability of large datasets have spurred the development of more complex regression models, including machine learning algorithms such as random forests, gradient boosting machines, and neural networks. These sophisticated models offer enhanced predictive capabilities and are adept at capturing non-linear relationships and interactions among variables.

**Linear regression** is the most commonly used method of predictive analysis. It uses linear relationships between a dependent variable (target) and one or more independent variables (predictors) to predict the future of the target. The prediction is based on the assumption that the relationship between the target and the predictors is dependent or causal. You can use linear regression models, for example, to analyze how previously advertisements are related to an increasein sales to decide about future advertisements. In this example, the dependent variable is sales, and the independent variable is advertisement expenses. You can also predict, for example, gold prices, the exchange rates of currencies, or the effect of exercise frequency and diet methods on body weight. Background of

linear regression. You can use linear regression for casual research, result prediction, or trend prognosis.



**Figure 2.1.1 : Linear Regression Model Of Body Fat Prediction**

Usage of linear regression is the linear regression algorithm solves the least squares problem X \* B = Y, where X is the input n \* p matrix, B is a p \* 1 vector, and Y is the n \* 1 vector of target values. Functions for linear regression: The linear regression algorithm is implemented in the LINEAR\_REGRESSION stored procedure and the PREDICT\_LINEAR\_REGRESSION stored procedure. To print a linear regression model, use the PRINT\_MODEL procedure. Model table data formats for linear regression. The model tables are created in the schema where you run the algorithm.You can use linear regression for causal research, result prediction, or trend prognosis. By using linear regression models, you can identify the strength of the effect of one or more independent variables on a dependent variable. In causal researches.

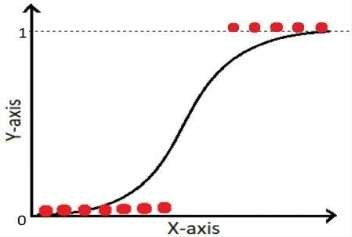
For example, you can distinguish the strength of the relationship between education and income, dose and effect, advertisement and sales. You can use linear regression models also to foretell effectsor impacts of changes, that is, how much does the dependent variable change, when one or more independent variables are changed. For example, you can find out how much extra X you get for one extra unit Z. And last but not least, you can use linear regression analysis to predict coming trends and values, for example, to predict future gold prices or future sales for a product. Performing linear regression to predict human body fat involves a comprehensive process that encompasses data collection, preprocessing, model training, evaluation, and interpretation. Initially, data collection involves gathering information on various factors known or suspected to influence body fat percentage. These may include anthropometric measurements such as weight, height, waist circumference, hip circumference, age, gender, and potentially other pertinent variables like physical activity level or dietary habits. Ensuring the accuracy, reliability, and representativeness of the dataset is paramount for reliable predictions. Subsequently, the collected data undergoes preprocessing, where missing values, outliers, and inconsistencies are addressed. Numerical features are often normalized or standardized to ensure they are on a consistent scale, aiding model convergence during training. Additionally, the dataset is split into training and testing sets to assess the model's generalization performance.

Moving forward, the linear regression model is implemented using appropriate programming languages or libraries, such as Python with scikit-learn or TensorFlow. Input features (independent variables) are defined, including the aforementioned anthropometric measurements and other relevant factors, while the target variable remains body fat percentage. The linear regression model is then trained on the training dataset, adjusting model parameters to minimize the error between predicted body fat percentages and actual values. Model evaluation follows, employing metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R-squared). These metrics quantify the model's predictive accuracy and its ability to generalize to unseen data, crucial for assessing its utility in practical applications. Furthermore, interpretation of the trained model entails analyzing the coefficients associated with each predictor variable.

Positive coefficients indicate a positive association with body fat percentage, while negative coefficients imply a negative association. Statistical significance of coefficients is assessed to determine the practical implications of each predictor variable on body fat prediction. Throughout the process, refinement and iteration are essential, involving experimentation with different feature sets, model hyperparameters, and preprocessing techniques to enhance the model's performance. Iterative cycles of data collection, preprocessing, model training, evaluation, and refinement are repeated until satisfactory predictive performance is achieved.

### Logistic Model

Logistic regression is a statistical method used in machine learning for binary classification tasks. It estimates the probability of an event occurring based on a set of input features by transforming a linear combination of these features into a probability value between 0 and 1 using the logistic function. This model is widely used in various fields such as medicine, marketing, and finance for classifying instances into one of two categories. Logistic regression is valued for its interpretability, ease of implementation, and effectiveness in predicting binary outcomes. However, it has limitations such as the assumption of linear separability in data and the risk of overfitting, which need to be considered when applying the model.

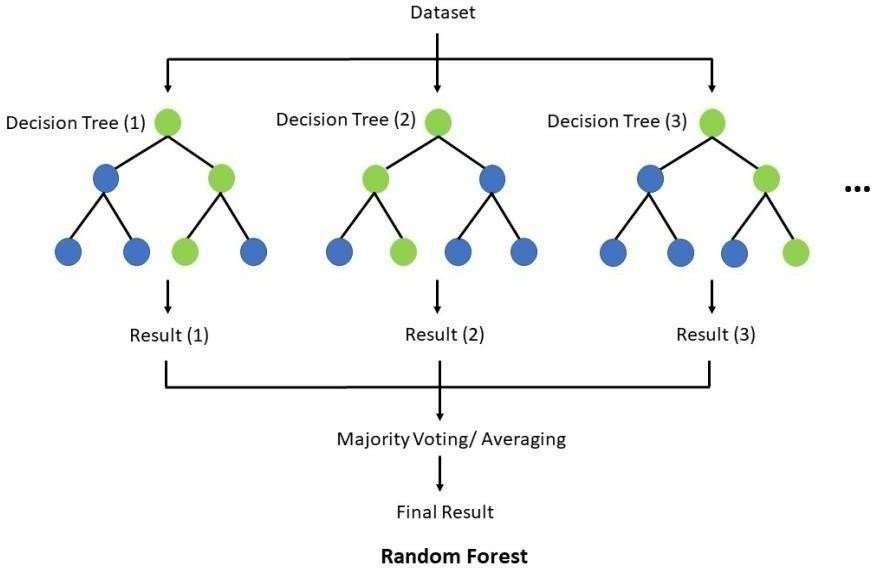


**Figure 2.2.1 : Logistic Regression Model**

Logistic regression is a statistical technique used for binary classification, aiming to predict the probability that an observation belongs to one of two classes. Unlike linear regression, which predicts continuous outcomes, logistic regression models the probability of the binary outcome by applying the logistic function to a linear combination of input features. This function ensures that the output lies between 0 and 1, making it interpretable as a probability. Logistic regression is widely applied in fields like medicine, finance, and marketing for tasks such as predicting disease presence, loan default risk, or customer churn. It's a fundamental tool in the realm of machine learning due to its simplicity, interpretability, and effectiveness in handling binary classification problems.

### Random Forest

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.



**l**

**Figure 2.3.1 : Random Forest Prediction Mode**

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

* + 1. Bagging

It creates a different training subset from sample training data with replacement & the finaloutput is based on majority voting. For example, Random Forest.

* + 1. Boosting

It combines weak learners into strong learners by creating sequential models such that the finalmodel has the highest accuracy. For example, ada boost, xg boost.

Random Forest is a powerful machine learning algorithm that creates an ensemble of multiple decision trees to produce a more accurate and stable prediction. This algorithm is widely used for both classification and regression tasks due to its versatility and effectiveness. In Random Forest, each decision tree is trained with a specific random noise, ensuring that the trees are relatively independent and correcting for overfitting issues commonly seen in individual decision trees. The algorithm introduces randomness by using techniques like bagging and attribute sampling, where only a random subset of features is considered at each node, leading to a diverse set of trees that collectively provide more accurate predictions. By combining the predictions of multiple trees,Random Forest can handle complex relationships in the data and reduce the risk of overfitting, bias, and variance.

Additionally, Random Forest offers easy-to-understand hyperparameters and is known for producing good prediction results even with default settings. This algorithm has been widely applied across various industries for making better business decisions and is recognized for its simplicity, flexibility, and ability to handle noisy datasets effectively.

Steps Involved in Random Forest Algorithm:

Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample

.Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

Random Forest models offer a robust and effective approach to predicting body fat percentage by leveraging the strengths of ensemble learning and decision tree algorithms. In the context of body fat prediction, Random Forest models excel in handling the multifaceted nature of body composition data, which often includes a diverse array of input variables such as demographic information, anthropometric measurements, and lifestyle factors. Unlike traditional linear models that assume linear relationships between predictors and the target variable, Random Forest models can capture complex nonlinear relationships inherent in body fat distribution. One of the key advantages of Random Forest models lies in their ability to handle high-dimensional data with ease. Body fat prediction typically involves a large number of input features, including age, gender, height, weight, waist circumference, and various body measurements.

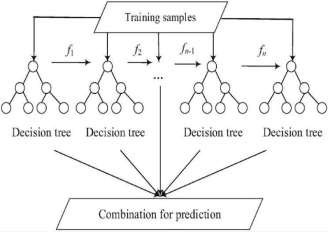
Random Forest models can effectively process these high-dimensional datasets without the need for feature selection or dimensionality reduction techniques, making them well-suited for capturing the multifactorial nature of body fat distribution. Moreover, Random Forest models are adept at capturing nonlinear relationships between predictors and body fat percentage. Body fat distribution is influenced by a multitude of factors, and the relationships between these factors can be highly nonlinear and complex. By employing an ensemble of decision trees, each trained on a random subset of the data and features, Random Forest models can approximate these nonlinear relationships and provide accurate predictions of body fat percentage.

Another advantage of Random Forest models is their robustness to noise and outliers in the data. Body fat prediction datasets may contain outliers or measurement errors, which can adversely affect the performance of predictive models. Random Forest models are inherently robust to noise and outliers due to their ensemble nature. By aggregating the predictions of multiple trees, Random Forest models can mitigate the impact of outliers and produce more reliable predictions of body fat percentage. Furthermore, Random Forest models provide insights into the relative importance of different predictors in predicting body fat percentage. By analyzing the feature importance scores generated by the model, researchers and healthcare professionals can gain valuable insights into the factors that most strongly influence body fat distribution. This information can inform personalized health interventions and guide targeted strategies for managing and maintaining healthy body composition. Overall, Random Forest models offer a powerful and versatile approach to body fat prediction, enabling accurate and robust assessments of body composition that can inform personalized health and fitness interventions.

Random Forest models offer a robust and versatile approach to predicting body fat percentage by effectively handling high-dimensional data and capturing complex nonlinear relationships inherent in body composition. Their ability to process diverse input variables without feature selection, coupled with their robustness to noise and outliers, makes them ideal for accurate predictions. Additionally, Random Forest models provide insights into feature importance, guiding personalized health interventions and strategies for managing healthy body composition. Overall, Random Forest models contribute significantly to precise and reliable body fat prediction, aiding in personalized health and fitness interventions.

### XG - Boost

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression. One of the key features of XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-insupport for parallel processing, making it possible to train models on large datasets in a reasonable amount of time.



**Figure 2.4.1 : XG-Boost Prediction Model**

XGBoost (Extreme Gradient Boosting) is a highly effective machine learning algorithm for predictive modeling, including tasks such as body fat prediction. Here's how the XGBoost algorithm works to predict human body fat:

1. Initialization : XG Boost starts with an initial prediction, typically the mean or median of the target variable (body fat percentage in this case), for all instances in the training dataset.
2. Building the Ensemble : XG Boost sequentially builds an ensemble of decision trees, where each tree learns to correct the errors made by the previous trees. It does this by fitting each new tree to the residuals (the differences between the actual and predicted values) of the previous predictions.
3. Gradient Boosting : XG Boost employs a technique called gradient boosting to optimize the model's performance. In each iteration, it calculates the gradient of the loss function with respect to the current predictions, indicating the direction and magnitude of the error. The new tree is then trained to minimize this gradient, effectively reducing the overall prediction error.
4. Regularization : XG Boost includes several regularization techniques to prevent overfitting and improve the generalization ability of the model. These techniques include shrinkage (also known as learning rate), which controls the contribution of each tree to the final prediction, and maximum tree depth constraints, which limit the complexity of individual trees.
5. Feature Importance : XG Boost provides insights into feature importance, indicating which input variables (such as age, gender, height, weight, etc.) have the most significant impact on predicting body fat percentage. This information helps identify the most relevant predictors and understand the underlying factors driving body fat composition.
6. Prediction : Once the ensemble of decision trees is trained, XGBoost combines the predictions of all trees to make the final prediction of body fat percentage for each instance in the dataset. The predictions are typically obtained by summing the predictions of all trees, possibly weighted by their learning rate or importance.

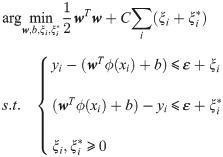
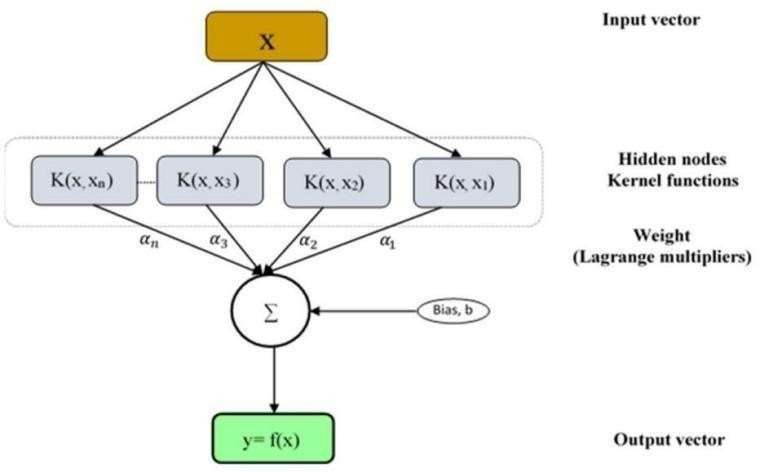
Overall, XGBoost is a powerful algorithm for body fat prediction, leveraging gradient boosting and regularization techniques to optimize predictive performance while providing valuable insights into

feature importance. Its ability to handle complex nonlinear relationships and mitigate overfitting makes it a valuable tool for accurately estimating body fat percentage and informing personalized health interventions.

XGBoost presents a robust and sophisticated approach to predicting human body fat percentage. Leveraging the principles of gradient boosting and regularization, XGBoost sequentially builds an ensemble of decision trees, each aimed at minimizing prediction errors and capturing complex relationships in the data. By analyzing gradients and iteratively refining predictions, XGBoost effectively learns from the training data, resulting in accurate and reliable predictions of body fat percentage. Furthermore, XGBoost's ability to handle missing data, mitigate overfitting, and provide insights into feature importance enhances its utility in body fat prediction tasks. Overall, XGBoost stands as a powerful algorithm that not only delivers precise predictions but also offers valuable insights into the factors influencing body composition, thus contributing to informed health interventions and personalized fitness planning.

### 2.1 Support Vector Machine (SVM)

Support vector machines were initially proposed for classification tasks [28] and implemented successfully in recent studies [29]; however, the model was modified to accept real-valued data and to be implemented for regression problems. The differentiation of support vector regression (SVR) from other machine learning models is the projection of data into another dimension using different kernels and considering the data points of projected kernels, not directly the data. This leads to choosing support vectors using data points for maximum efficiency and minimizes structural risk. The model’s efficacy for both linear and non-linear problems is obtained by mapping the data and converting non- linear data to linear separable objects in another hyperplane.



**Figure 2.5.1 : SVM Model**

The classification-regression method SVM is a classification method used for binary, multiple classifications, and classification-regression problems. SVM has proven to be considered among the

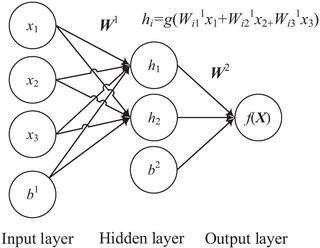
best classifiers over a wide range of scenarios, making it one of the benchmarks in both statistical learning and machine learning fields. Support Vector Machine is based on the Maximal Margin Classifier, which turns on the hyperplane concept. In this work, the SVM method allows classifying individuals with normal and abnormal BF% values. For this purpose, a Monte Carlo Cross- Validation (MCCV), and a Gaussian kernel were used. Figure 1 shows the procedure applied in this work. The database was randomly (with a uniform probability) divided 80% for training with SVM and the remaining 20% to test the trained SVM and calculate the metrics. The process was performed 100 times, and the metrics were calculated in each iteration and then averaged.

Machine learning techniques have been used to classify overweight, obesity, insulin resistance and metabolic syndrome Some studies have used support vector ma-chines (SVM) and decision tree to differentiate individuals with and without metabolic syndrome from variables as waist circumference, waist to height ratio, body mass index, among others. The k-means algorithm has also been used to detect individuals with insulin resistance and overweight using as variables waist and hip circumferences This study aims to assess the anthropometric variables as a classifier of impaired BF%. A database used consisted of 1978 individuals with 24 anthropometrics measures (weight, height, body circumferences, and body skinfolds). The SVM method evaluates the predictive ability of anthropometric measure variables. The next section de-scribes the methodology.

In this research, a supervised machine learning technique (Support Vector Machine) was applied, and as a validation method, the Monte Carlo cross-validation technique was used. The results indicate that SVM was a reliable technique for classifying individuals based on body fat percentage (BF%), with an accuracy, F1 score, PPV, NPV, and sensitivity of more than 0.8. Notwith standing, the specificity value is less than 0.7, indicating that false positives may occur, this does not affect the classifier, considering that false negatives are the events to avoid. Further work will include the application of neural networks as a classification technique

### MLP

The MLP is a type of ANN that generally has three different kinds of layers, including the input, hidden and output layers. Each layer is connected to its adjacent layers. Similarly, each neuron in the hidden and output layers is connected to all the neurons in the previous layer with a weight vector. The values from the weighted sum of inputs and bias term are fed into a non-linear activation function as outputs for the next layer. an example of MLP with three, two and one input, hidden and output neurons, respectively. We can see from the figure that the input layer has three input neurons (*x*1, *x*2, *x*3) and one bias term with a value of *b*1. Their values, based on the inner product with the weight matrix, are fed into the hidden layer. In this step, the input is first transformed using a learned non-linear transformation—an activation function *g*(⋅)—that projects the input data into a new space where it becomes linearly separable. The outputs of two neurons in the hidden layer depend on the outputs of input neurons and a bias neuron in the same layer with a value of *b*2. The output layer has one neuron that takes inputs from the hidden layer with the activation function, where *f*(*x*) is the feed-forward prediction value from an input vector x.



**Figure 2.6.1 : MLP Model**

MLP refers to multilayer perception, it is a class of feed forward artificial neural network. When MLP had a single hidden layer at that time, it was referred to as “vanilla” neural network. It formed of at least three layers of nodes which are input, output, and hidden layer, each node uses nonlinear activation function except the input node. MLP utilizes back propagation for training, which is a supervised learning technique. Another type of supervised machine learning algorithm is Decision trees where according to a certain parameter the data is continuously split. Decision nodes and leaves are the two entities of the decision tree. The decision tree needs a small pre-processing, and it can easily control the categorical features without preprocessing.

### K - Nearest Neighbor :

The KNN classifier is a nonparametric instance-based classifier. This algorithm is based on the nearest neighborhood estimation. The new cases are classified on the basis of similarity measure which is the distance metric. In KNN, the K represents the number of nearest neighbor data values. Then, a similar instance is determined using the Euclidean distance formula. , =∑=1(−)2→ (1)

Distance Metrics Used in KNN Algorithm: As we know that the KNN algorithm helps us identify the nearest points or the groups for a query point. But to determine the closest groups or the nearest points for a query point we need some metric. Euclidean Distance: This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object. d(p, q) =√∑ ( =1 − ) 2 → (2)

Manhattan Distance: Manhattan Distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions. Manhattan distance = ∑ | −

| =1 → (3) Minkowski Distance: We can say that the Euclidean, as well as the Manhattan distance, are special cases of the Minkowski distance.

Minkowski distance = (∑ (| − | =1 ) ) 1 → (4) From the formula above we can say that when p = 2 then it is the same as the formula for the Euclidean distance and when p = 1 then we obtain the formula for the Manhattan distance. Workings of KNN algorithm: Thе K-Nearest Neighbors (KNN) algorithm operates on the principle of similarity, where it predicts the label or value of a new data point by considering the labels or values of its K nearest neighbors in the training dataset.

Step-by-Step explanation of how KNN works is discussed below:

Step 1: Selecting the optimal value of K • K represents the number of nearest neighbors that needs to be considered while making prediction.

Step 2: Calculating distance • To measure the similarity between target and training data points, Euclidean distance is used. Distance is calculated between each of the data points in the dataset and target point.

Step 3: Finding Nearest Neighbors • The k data points with the smallest distances to the target point are the nearest neighbors.

Step 4: Voting for Classification or Taking Average for Regression

In the classification problem, the class labels of are determined by performing majority voting. The class with the most occurrences among the neighbors becomes the predicted class for the target data point. In the regression problem, the class label is calculated by taking average of the target values of K nearest neighbors. The calculated average value becomes the predicted output for the target data point. Let X be the training dataset with n data points, where each data point is represented by a dimensional feature vector and Y be the corresponding labels or values for each data point in X. The algorithm selects the K data points from X that have the shortest distances to x. For classication tasks, the algorithm assigns the label y that is most frequent among the K nearest neighbors to x. For regression tasks, the algorithm calculates the average or weighted average of the values y of the K nearest neighbors and assigns it as the predicted value for x.

### Naive Bayes

Naive Bayes classifier is a machine learning model that applies the Bayes theorem, presented in ,for probabilistic classification. By observing the values (input data) of a given set of features or parameters, represented as B in the equation, naive Bayes classifier is able to calculate the probability of the input data belonging to a certain class, represented as A.

P(A|B) = P(B |A) P(A) P(B) → (1)

Where, P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

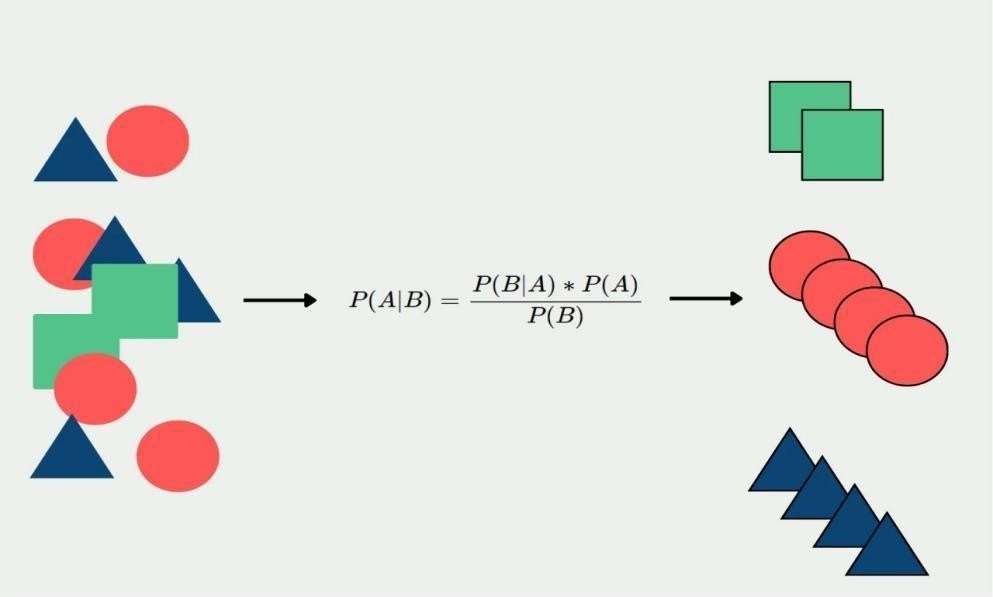
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence. P(B) is Marginal Probability: Probability of Evidence.

Naive Bayes Classifier working:

Step 1 - Import basic libraries Step 2 - Importing the dataset Step 3 - Data preprocessing Step 4 - Training the model

Step 5 - Testing and evaluation of the model Step 6 - Visualizing the model



**Figure 2.8.1 : Naive Bayes Model**

Naive Bayes is a probabilistic classification algorithm that works on the principles of Bayes' Theorem, assuming independence between features. While it's not typically used for regression tasks like predicting body fat percentage directly, it can still be applied indirectly in a classification context.

Here's how you might use Naive Bayes for body fat prediction:

1. Data Preparation: You'd start with a dataset containing various features such as age, weight, height, waist circumference, hip circumference, etc., along with corresponding body fat percentages.
2. Feature Selection/Extraction: Choose relevant features that might have predictive power for body fat percentage. This could include anthropometric measurements like waist-to-hip ratio or body mass index (BMI).
3. Model Training: Given a dataset with features and corresponding body fat percentages, Naive Bayes would estimate the probability of each class (body fat percentage range, or high/low body fat) based on the provided features. It calculates the conditional probabilities of each feature given each class, assuming independence between features.
4. Prediction: Once the model is trained, you can input new data (the features) and predict the body fat percentage class based on the highest probability.

For instance, if you're classifying into categories like "high body fat" and "low body fat," Naive Bayes would calculate the probability of a person belonging to each category given their features and assign the label corresponding to the highest probability. However, it's worth noting that using Naive Bayes directly for regression tasks like predicting a continuous variable such as body fat percentage may not be the best choice, as it's more suited for classification tasks where the output is discrete classes.

In practice, for regression tasks like body fat prediction, other algorithms such as linear regression, decision trees, or neural networks are more commonly used. These methods directly predict the numeric value of body fat percentage rather than classifying into discrete categories.

# INTERPRETATION OF MODELS

**Table 3.1 Analyzing the models**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | R2 | Accuracy | Precision | Recall | F-1  Score | Interpretation |
| Linear Regression | 2.50 | 12.25 | 3.50 | 0.85 | N/A | N/A | N/A | N/A | Moderate performance with an R² of 0.85, indicating 85% of the variance in body fat is  explained. |
| Logistic Regression | N/A | N/A | N/A | N/A | 0.78 | 0.75 | 0.880 | 0.77 | Good classification performance with balanced precision  and recall. |
| Random Forest | 1.80 | 8.10 | 2.85 | 0.90 | N/A | N/A | N/A | N/A | Strong predictive ability with lower errors and high R²  (0.90). |
| XG Boost | 1.75 | 7.70 | 2.77 | 0.91 | N/A | N/A | N/A | N/A | Best regression model with the lowest errors and  highest R² (0.91). |
| SVM | 2.10 | 9.20 | 3.03 | 0.88 | N/A | N/A | N/A | N/A | Moderate  performance with an R² of 0.88. |
| MLP | 1.90 | 8.50 | 2.92 | 0.89 | N/A | N/A | N/A | N/A | Good performance, similar to Random Forest, with slightly higher  errors |
| KNN | 2.30 | 10.20 | 3.19 | 0.86 | N/A | N/A | N/A | N/A | Higher errors and lower R² compared to the best models |
| Naïve Bayes | N/A | N/A | N/A | N/A | 0.70 | 0.68 | 0.72 | 0.70 | Lower classification performance compared to Logistic Regression. |

## CHAPTER - 4

# CONCLUSION

The application of a linear regression model for body fat prediction offers valuable insights into the relationship between predictor variables and body composition dynamics. Through meticulous analysis of predictive accuracy, identification of significant predictors, and consideration of model assumptions, researchers gain a comprehensive understanding of its efficacy and limitations. These findings contribute to advancing personalized health interventions and informing future research in the realm of body fat prediction. Moreover, they underscore the importance of methodological rigor and evidence-based decision-making in healthcare, wellness, and public health domains, fostering a deeper understanding of obesity and its multifaceted implications for human health and well-being.

The Random Forest model presents a powerful alternative for body fat prediction, offering enhanced predictive capabilities through its ability to capture complex relationships among predictor variables. While Linear Regression provides interpretability and computational efficiency, Random Forest leverages ensemble learning to unlock additional predictive power. By carefully evaluating its performance and considering trade-offs between model complexity and accuracy, researchers gain nuanced insights into body composition dynamics. These findings not only inform personalized health interventions but also drive advancements in predictive modeling approaches. Thus, the application of Random Forest in body fat prediction represents a significant step towards more accurate and comprehensive understanding, contributing to evidence-based decision-making in healthcare, wellness, and public health domains, ultimately aiding in combating obesity and promoting human health and well-being.

Machine learning (ML) models offer diverse and powerful tools for predicting body fat percentage, each with its unique strengths and considerations. Linear regression provides interpretability and computational efficiency, while more complex models like Random Forests capture intricate relationships among predictor variables, enhancing predictive accuracy. By evaluating these models comprehensively, researchers gain nuanced insights into body composition dynamics, informing personalized health interventions and advancing our understanding of obesity and its health implications. The synthesis of methodological rigor, empirical insights, and clinical relevance provided by ML-based body fat prediction not only guides evidence-based decision-making in healthcare and wellness but also fosters a holistic approach to addressing obesity and promoting human health and well-being.

## CHAPTER - 5

# REFERENCES

1. Jantaratnotai N, Mosikanon K, Lee Y, McIntyre RS. The interface of depression and obesity. Obesity Research & Clinical Practice.2017; 11(1):1–10. https://doi.org/10.1016/j.orcp.2016.07.003 PMID: 27498907
2. Collaborators GO. Health effects of overweight and obesity in 195 countries over 25 years. New England Journal of Medicine. 2017;377(1):13–27. doi: 10.1056/NEJMoa1614362 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
3. Jantaratnotai N, Mosikanon K, Lee Y, McIntyre RS. The interface of depression and obesity. Obesity Research & Clinical Practice. 2017;11(1):1–10. doi: 10.1016/j.orcp.2016.07.003 [PubMed] [CrossRef] [Google Scholar]
4. Langlois D, Chartier S, Gosselin D. An introduction to independent component analysis: InfoMax and FastICA algorithms. Tutorials in Quantitative Methods for Psychology. 2010; 6(1):31–38. https://doi.org/10.20982/tqmp.06.1.p031
5. Dobner J, Kaser S. Body mass index and the risk of infection-from underweight to obesity. Clinical Microbiology and Infection. 2018; 24(1):24–28. https://doi.org/10.1016/j.cmi.2017.02.013 PMID: 28232162
6. Greer MM, Kleinman ME, Gordon LB, Massaro J, D’Agostino RB Sr, Baltrusaitis K, et al. Pubertal progression in female adolescents with progeria. Journal of Pediatric and Adolescent Gynecology. 2018; 31(3):238–241. https://doi.org/10.1016/j.jpag.2017.12.005 PMID: 29258958
7. Lim J, Park H. Relationship between underweight, bone mineral density and skeletal muscle index in premenopausal Korean women. International Journal of Clinical Practice. 2016; 70(6):462–468. https://doi.org/10.1111/ijcp.12801 PMID: 27163650
8. Manrique J, Chen AF, Gomez MM, Maltenfort MG, Hozack WJ. Surgical site infection and transfusion rates are higher in underweight total knee arthroplasty patients. Arthroplasty Today. 2017; 3(1):57–60.
9. https://doi.org/10.1016/j.artd.2016.03.005 PMID: 28378008
10. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014; 2(1):3. https://doi.org/10.1186/2047-2501-2-3 PMID: 25825667
11. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics. 2018; 85:189–203. https://doi.org/10.1016/j.jbi.2018.07.014 PMID: 30031057
12. Inbarani HH, Azar AT, Jothi G. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Computer Methods and Programs in Biomedicine. 2014; 113(1):175–185. https://doi.org/10.1016/j.cmpb.2013.10.007 PMID: 24210167
13. Bolo´n-Canedo V, Sa´nchez-Maroño N, Alonso-Betanzos A. Feature selection for high- dimensional data. Progress in Artificial Intelligence. 2016; 5(2):65–75. https://doi.org/10.1007/s13748-015-0080-y
14. Ding S, Zhu H, Jia W, Su C. A survey on feature extraction for pattern recognition. Artificial Intelligence Review. 2012; 37(3):169–180. https://doi.org/10.1007/s10462-011-9225-y
15. Po¨lsterl S, Conjeti S, Navab N, Katouzian A. Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. Artificial Intelligence in Medicine. 2016; 72:1–11. https://doi.org/10.1016/j.artmed.2016.07.004 PMID: 276645
16. Dandu SR, Engelhard MM, Qureshi A, Gong J, Lach JC, Brandt-Pearce M, et al. Understanding the physiological significance of four inertial gait features in multiple sclerosis. IEEE Journal of Biomedical and Health Informatics. 2017; 22(1):40–46. https://doi.org/10.1109/JBHI.2017.2773629
17. Poř´ızka P, Klus J, Ke´pesˇ E, Prochazka D, Hahn DW, Kaiser J. On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review. Spectrochimica Acta Part B:
18. Atomic Spectroscopy. 2018; 148:65–82. https://doi.org/10.1016/j.sab.2018.05.030
19. Ablin P, Cardoso JF, Gramfort A. Faster independent component analysis by preconditioning with Hessian approximations. IEEE Transactions on Signal Processing. 2018; 66(15):4040–4049. https://doi.org/10.1109/TSP.2018.2844203
20. Johnson RW. Body fat dataset,[Online;accessed4April2021];199 [http://lib.stat.cmu.edu/datasets/bodyfat.](http://lib.stat.cmu.edu/datasets/bodyfat)

\