Exploring Body Fat Composition: A Machine Learning Approach

Khush Naidu, *San Jose State University, Computer Science,* Breanna Chi, *San Jose State University, Computer Science.*

***Abstract*— This project delves into the intricate relationship between body fat percentage and various anthropometric measurements, such as weight, age, and specific body part dimensions. Divided into three distinct parts, the study employs machine learning techniques to extract valuable insights from a comprehensive dataset.**

***Keywords*— body fat percentage, machine learning, anthropometric measurements, body type analysis, age-related analysis.**

I. Introduction

The study of body composition and its implications for health and well-being has gained significant traction in recent years. Advances in machine learning present an opportunity to delve deeper into the complex interplay between body fat percentage and various anthropometric measurements. This project embarks on a comprehensive exploration, utilizing a diverse dataset containing weight, age, and specific body part dimensions. Divided into three distinct parts, our research employs cutting-edge machine learning techniques to extract nuanced insights into body composition dynamics.

II. Data

This project uses the Body Fat Prediction Dataset, which uses the estimations of body fat percentages for 252 men through underwater weighing and various body circumference measurements. The objective is to enhance the comprehension of multiple regression techniques, offering viable alternatives to inconvenient or expensive methods for precise body fat measurement. Specifically crafted for instructive purposes in multiple regression, the dataset facilitates the exploration of predicting body fat using easily accessible measurements, addressing the impracticality of accurate body fat measurement methods. Key variables in the dataset include body fat percentage derived from Siri's equation, age, weight, height, and diverse body circumference measurements, all adhering to Benhke and Wilmore's (1974) measurement standards. Siri's equation, rooted in body density, serves as a pivotal formula for estimating body fat percentage by considering the proportion of lean body tissue and fat tissue. The derivation of body density involves underwater weighing and the calculation of the weight difference in air and water. The dataset's significance lies in its pivotal role in the development of predictive equations for lean body weight, as exemplified by Penrose et al. (1985). These equations offer practical alternatives for estimating body fat, proving especially beneficial for health assessments where precise measurements pose challenges.

III. Body Fat Percent Prediction

The project utilizes a dataset containing body measurements, age, and weight as features, with body fat percentage as the target variable. This section is organized into three main subsections: Body Fat Percentage Prediction, Correlation Mapping, and training and Testing Our Model.

A. Feature Selection:

To achieve accurate predictions, two feature selection techniques, Recursive Feature Elimination (RFE) and Decision Tree Regressor, were employed. The RFE method ranked features based on their importance in predicting body fat percentage, while the Decision Tree Regressor assessed feature importances. The cumulative ranking system was introduced to reconcile results from both techniques. The final selected features are Abdomen, Weight, Wrist, Thigh, and Hip.

B. Correlation Mapping:

A heatmap analysis was conducted to explore correlations between features. Highly correlated features were identified and a threshold-based function was applied to narrow down the selection. The final set of features for body fat prediction includes Weight, Abdomen, Thigh, and Wrist.

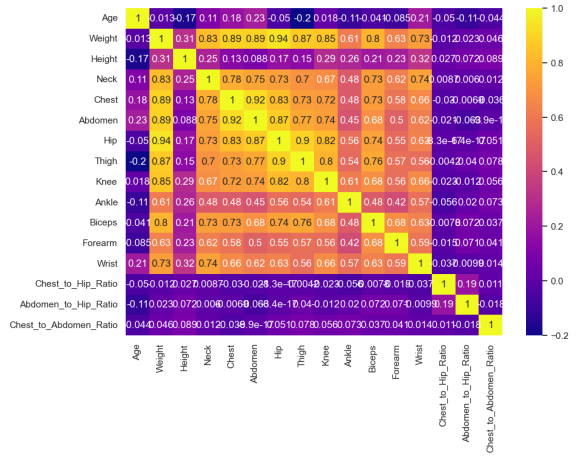

Fig. 1 Correlation Map for all features in the dataset to identify closely related features.

C. Highly Correlated Features:

Thigh and Hip were identified as similar, as were Chest and Abdomen. To avoid redundancy, one feature from each pair was selected. The final list of features is Weight, Abdomen, Thigh, and Wrist.

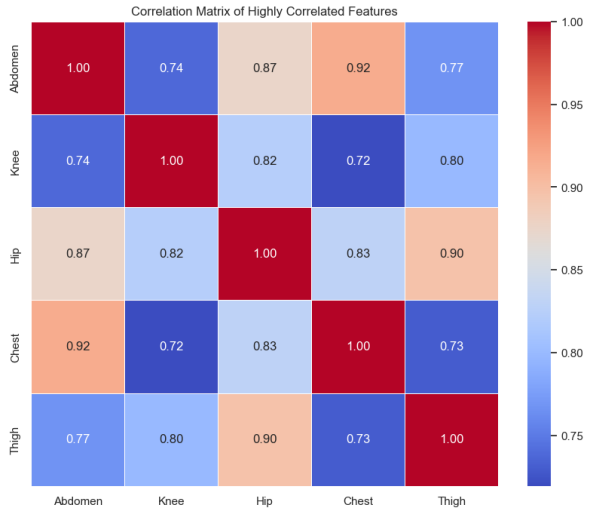

Fig. 2 A correlation map between the selected highly correlated features.

D. Training and Testing Our Model:

Three classifiers—Support Vector Machine (SVM), Random Forest Classifier, and K-Nearest Neighbor—were employed to train and test the model based on the selected features. Here are the results for each model:

1. SVM Classifier:

Training Accuracy: 0.76
Training F1 Score: 0.76
Test Accuracy: 0.61
Test F1 Score: 0.58

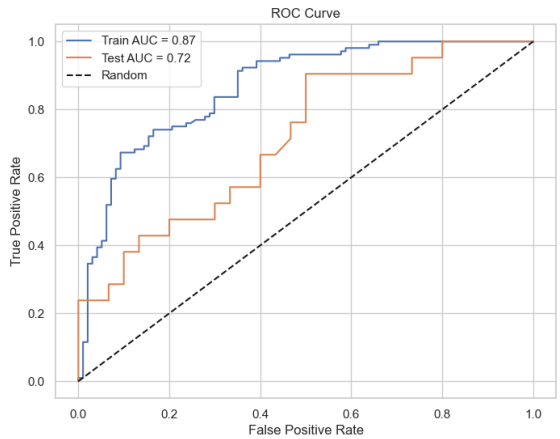

Fig. 3 ROC Curve for SVM Classifier

2. Random Forest Classifier:

Training Accuracy: 0.80
Training F1 Score: 0.81
Test Accuracy: 0.75
Test F1 Score: 0.72

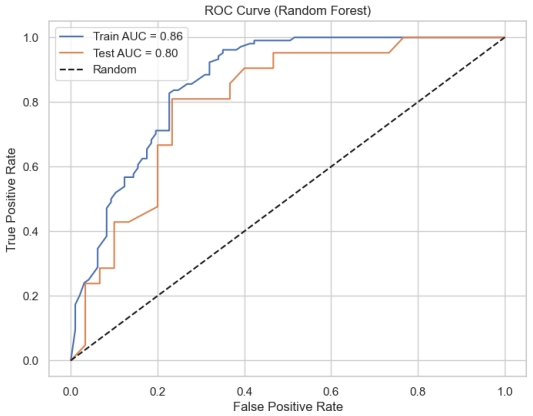

Fig. 4 ROC Curve for Random Forest Classifier

3. K-Nearest Neighbor Classifier:

Training Accuracy: 0.80
Training F1 Score: 0.81
Test Accuracy: 0.73
Test F1 Score: 0.71

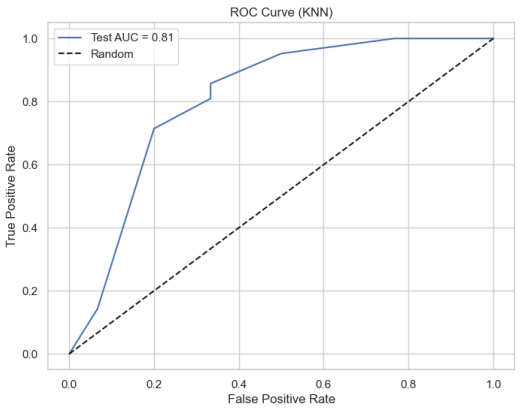

Fig. 5 ROC Curve for KNN Classifier

E. Model Comparison and Selection

The Random Forest Classifier stands out as the most effective model for predicting body fat percentage. Its robust evaluation, including ROC analysis and confusion matrix visualization, supports its superior performance.

IV. Body Type Analysis

In this phase, the project aims to identify common body types within the male population using engineered features. By employing body measurement ratio features, specifically

Chest to Hip, Abdomen to Hip and Chest to Abdomen ratios. Cluster analysis is performed to unveil potential trends. The chosen clustering algorithm is K-means, recognized for grouping similar data points based on specified features.

A. Feature Engineering:

Features of interest for clustering were engineered using the initial dataset. They are as follows: 'Chest_to_Hip_Ratio,' 'Abdomen_to_Hip_Ratio,' 'Chest_to_Abdomen_Ratio.'

B. Data Scaling:

MinMaxScaler was applied to scale the data, ensuring uniformity.

C. Determining Optimal Clusters:

The Elbow Method was utilized to identify the optimal number of clusters (k).

D. K-means Clustering:

Optimal k (number of clusters) was chosen based on the Elbow Method. K-means clustering was applied to categorize data points into clusters.

E. Visualization in 3D Space:

Clusters were visualized in a 3D plot for a comprehensive understanding. Axes represent normalized ratios: Chest-to-Hip, Abdomen-to-Hip, Chest-to-Abdomen.

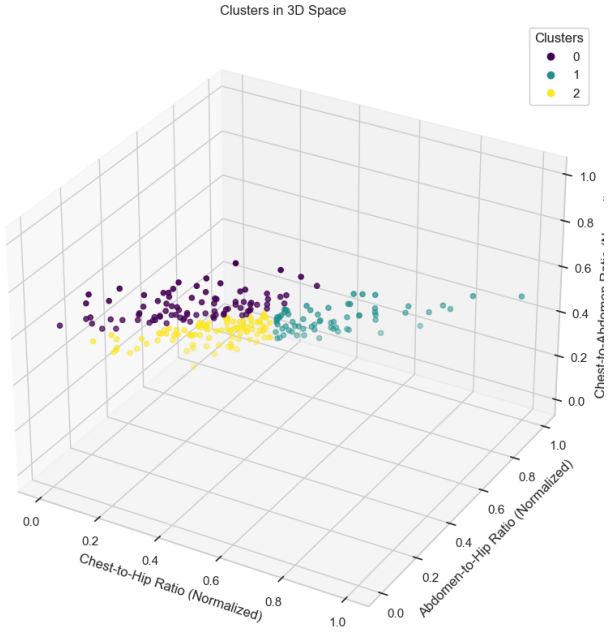

Fig. 5. 3D Visualization of Clusters

F. Results:

The number of optimal clusters was determined to be 3. The following table records the mean values for each ratio pertaining to each cluster:

Displaying the mean values for each cluster across the selected features.

Cluster	Chest to Hip Ratio	Abdomen to Hip Ratio	Chest to Abdomen Ratio
1	0.999621	0.867453	1.152887
2	1.060128	0.991943	1.069646
3	0.983144	0.929132	1.058797

Table.1 Mean Values for ratios pertaining to each cluster.

Identification of Clusters:

1. Cluster 1: Larger Hips compared to chest and abdomen (Gynoid or Pear-shaped body type).
2. Cluster 2: Larger Chest and similar-sized abdomen compared to hips (Trapezoid or Torch body type).
3. Cluster 3: Larger hips compared to chest and abdomen but larger chest compared to abdomen (Hourglass body type).

Hence, three body type trends were identified in the dataset population of men.

V. Aging and its Implications on Body Fat

The analysis delved into understanding the intricate relationship between aging and overall body fat. The objective was to discern whether aging predisposes individuals to gain or lose fat. Various regression models were employed to capture the nuanced trends in the dataset.

A. Feature Selection:

The features Age and target Body Fat were selected to study using linear regression.

B. Regression Models:

We decided to train four different regression models to compare results and find the best one suited for this regression problem. We used the metrics of Mean Squared Error (MSE) and Mean Absolute Error (MAE) to evaluate their performance:

Metric	Random Forest	Polynomial Regression	Support Vector	Gradient Boosting

	Regressor		Regression	Regressor
MSE	339.3535	316.4429	379.7285	332.6567
MAE	17.9422	17.6051	19.1782	17.8641

Table.2 MSE and MAE values for all four regression models

C. *Visualization of Model Performance:*

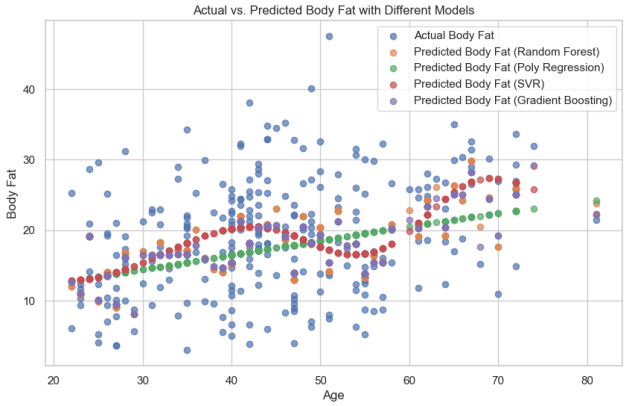

Fig. 8. Predictions for bodyfat with age fluctuation.

D. *Feature Engineering:*

To further understand the trends of body fat within different age groups, we decided to engineer another feature– Body Fat to Age Ratio – and study it using cluster analysis.

E. *K-Means Clustering:*

K-means clustering with three clusters was applied to the filtered data. Here is a visual of the clusters obtained:

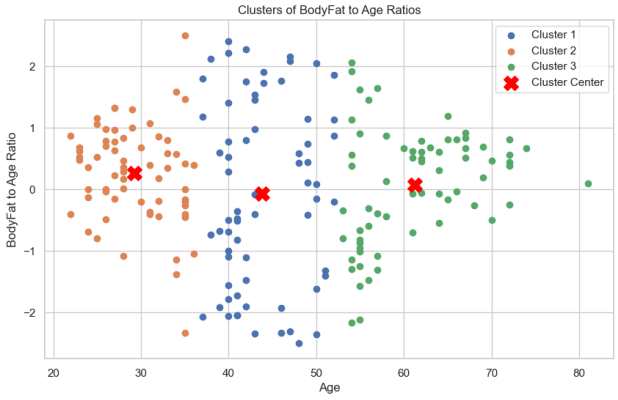

Fig. 8. Clusters of BodyFat-Age Ratios in the Dataset Population

F. *Results:*

Mean age and mean bodyfat_age_ratio were calculated for each cluster to better map and separate the results. Here are the cluster statistics:

Cluster	Mean Age	Mean Body Fat to Age Ratio
1	43.835821	-0.067847
2	29.230769	0.270831
3	61.238806	0.080098

Table.3 Different Body Fat To Age Ratio Trends.

From this, we were able to identify body fat trends in three different age groups. Here are the trends:

1. ***Cluster 1***, representing a younger age group, exhibits a higher mean Bodyfat_Age_Ratio, suggesting relatively higher body fat levels compared to the mean for that age.
2. ***Cluster 2,*** corresponding to an older age group, shows a slightly lower mean Bodyfat_Age_Ratio, indicating a potential decrease in body fat levels.
3. ***Cluster 3*** demonstrates a negative mean Bodyfat_Age_Ratio, suggesting a deviation from the general trend observed in the other clusters.

VI. Conclusion

In summary, this project employed machine learning techniques to explore the complex relationship between body fat percentage and anthropometric measurements. Divided into three segments, it successfully predicted body fat using the Random Forest Classifier, identified common male body types through K-means clustering, and investigated the impact of aging on body fat using regression models. The Random Forest Regressor and Polynomial Regression emerged as effective predictors for body fat levels across different ages. The analysis not only provided valuable insights into distinct body types prevalent in the male population but also revealed nuanced trends in body fat distribution concerning age. This comprehensive approach contributes to advancing the understanding of body composition dynamics and establishes a robust framework for future studies in personalized health assessments and fitness considerations.

VII. Future Work

In the near future, we are looking forward to creating a webapp prototype to implement our outcomes from this project.

VIII. Acknowledgment

The authors would like to thank Dr. Saptarshi Sengupta for his guidance and support on this project.

IX. References

[1] Fedesoriano. “Body Fat Prediction Dataset.” Kaggle, 14 June 2021, www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset.

[2] “Body Fat Distribution.” Body Fat Distribution - an Overview | ScienceDirect Topics, www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/body-fat-distribution#:~:text=Body%20fat%20distribution%20is%20a,absolute%20amount%20of%20body%20fat. Accessed 6 Dec. 2023.

[3] “Study Deflates Notion That Pear-Shaped Bodies More Healthy than Apples: Abnormal Proteins from Buttock Fat Linked to Metabolic Syndrome.” ScienceDaily, ScienceDaily, 10 Jan. 2013.

[4] Kousar, Sahar, and et. al. “Classification of Male Upper Body Shape: An Innovative Approach.” *Sage Journals Home*, journals.sagepub.com/doi/full/10.1177/15589250231177447. Accessed 11 Dec. 2023.

[5]Watson, Stephanie. “Waist-to-Hip Ratio: Chart, Ways to Calculate, and More.” *Healthline*, Healthline Media, 2 Feb. 2023, www.healthline.com/health/waist-to-hip-ratio.