Body Fat Prediction

Advanced Computer Science Masters Project

Abdul Jaleel Mohammed
22090668
Maria Psarrou

Section 1: Introduction and Overview

Obesity is defined as having an excessive amount of body fat. In the UK, it is estimated that approximately one in four adults and one in five children aged 10 to 11 are affected by obesity. The Body Mass Index (BMI) is the most commonly used tool for identifying obesity, with a BMI range of 30 to 39.9 indicating obesity. However, for individuals of Asian, Chinese, Middle Eastern, Black African, or African-Caribbean descent, a BMI of 27.5 or higher is used to classify overweight and obesity. Obesity can significantly increase the risk of developing various serious health conditions, including high blood pressure, asthma, and cancer (NHS, 2023).

The human body consists of extracellular fluid, bone, fat, and muscle cells. When these components are proportionally balanced, the body maintains optimal composition. White fat tissue, which accounts for about 15-20% of the body, is found in two main areas: under the skin and around internal organs. In infants, brown fat tissue makes up around 4% of body weight. Fat tissue is also considered an endocrine organ with the ability to regulate heat and energy, as well as having secretory functions. Excess energy is stored in white fat for future use. The primary fat cells, called adipocytes, are supported by fibroblasts, preadipocytes, and macrophages. For individuals with obesity, the key focus of treatment is the reduction of excess fat tissue (Uçar *et al.*, 2021, vol. 167). In healthy individuals, body fat percentages should fall between 25-30% for women and 18-23% for men. Women with more than 30% body fat and men with more than 25% are classified as obese (Penn Medicine, no date).

This project implements a unique approach to predicting obesity by creating machine-learning models that estimate body fat percentage using physical measurements such as weight, abdomen, chest, hip circumferences, and density determined from underwater weighing.

1.2 Research Question:

"Can machine learning models accurately predict body fat percentage in humans with multiple anthropometric measurements and density determined from underwater weighing to represent they are obese?"

1.3 Practical Investigation:

The investigation will involve several technical tasks:

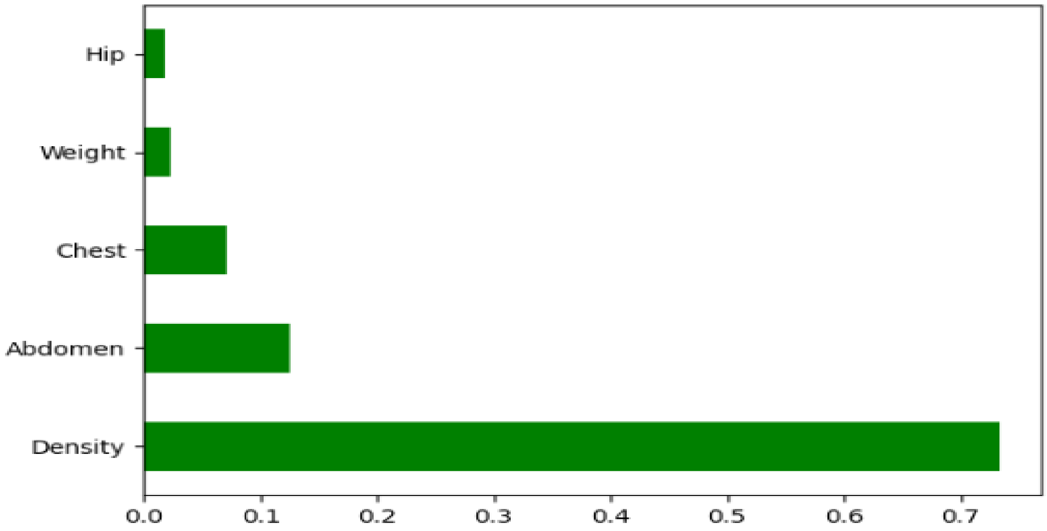
- **Dataset Construction**: In this study, a dataset was used for estimating body fat percentage from the online open-source platform. The data set includes anthropometric measurements and body fat percentage values of 252 individuals. It consists of 252 rows and 15 columns. The goal here is to calculate the Y Body Fat Percentage using the minimum X variable. Body fat percentage measurements were obtained using Siri Equations.

- **Tools and Techniques:** The programming language used for feature engineering, data management, and model construction was Python. Seaborn supported visualization, and libraries like Pandas and NumPy were crucial for data manipulation. To assess and choose the most informative features in relation to the target variable, a mutual information gain method was applied. By picking features with high predictive power, our technique made sure the model only used the most pertinent measurements. Utilized the feature_importance technique from the ExtraTreesRegressor class in a feature selection procedure. Concentrating on features with the highest scores, this method improved the model. To test for multicollinearity, the variance inflation factor was computed for every feature.
- **Model Development**: Various machine-learning models are utilized to analyze the relationship between these anthropometric measurements and body fat percentage. Three different Machine Learning models were used in this approach Support Vector Regression, RandomForestRegression, and DecisionTreeRgression are utilized to predict the Body fat percentage in individuals. Hyperparameter tuning was carried out to enhance the performance of the model.

1.4 Deliverables:

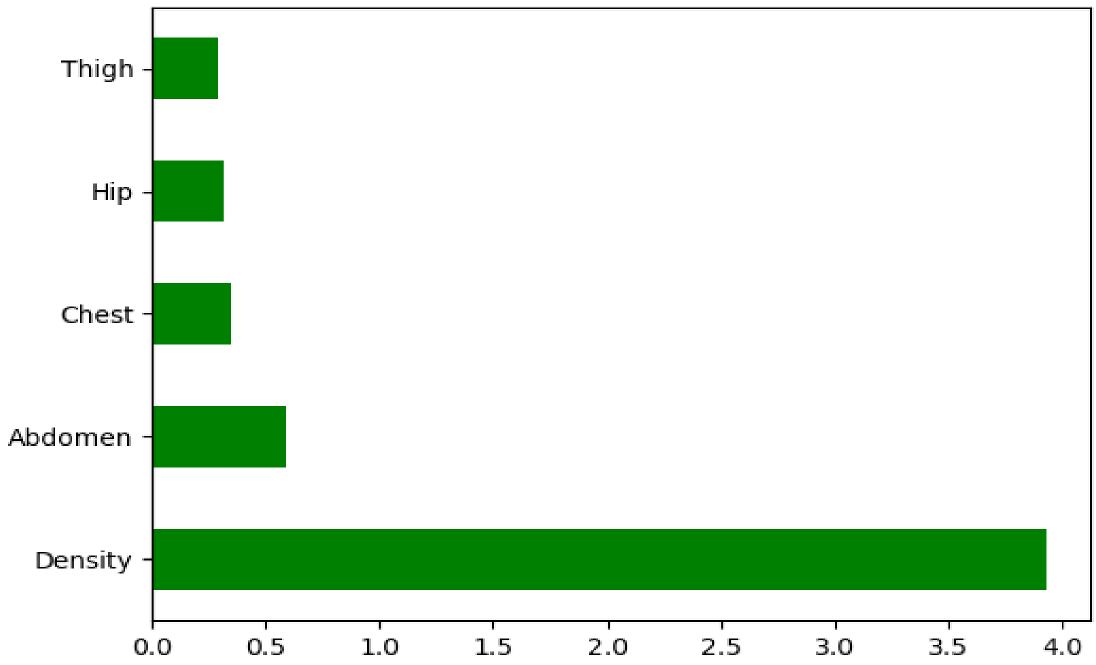
- **Data Preprocessing**:

A Python function using scipy.stats, matplotlib.pyplot, and seaborn libraries to plot the distribution of data columns. It includes histograms, regression plots, and box plots for each feature, making it easier to inspect and identify outliers across the dataset visually. The code utilizes the ExtraTreesRegressor to assess and rank feature importance, highlighting the top five most impactful features for predicting body fat percentage. This selection process refines the model by concentrating on the features that best contribute to predictive accuracy.



Fig(a): The Bar graph obtained from ExtraTreesRegressor Algorithm.

From Figure (a), The y-axis represents density, abdomen, weight, chest, and hip circumferences. The x-axis represents values with feature importances considered the five most impactful features on the target variable. The density determined from underwater weighing is greater than 0.7, and the Abdomen has a value greater than 0.1, Hip and weight have values less than 0.1.



Fig(b): Bar graph obtained from mutual information gain technique.

In Fig (b), The y-axis represents the top five features from the dataset. On the x-axis values are obtained from the mutual information gain in relation to the top 5 features. Density stands out as the most influential feature, with an importance value nearing approximately 4.0, highlighting its strong role in predicting body fat percentage.

The abdomen ranks as the second most significant feature, with an importance slightly above 0.5. Chest has a moderate impact, with an importance of just over 1.0. Hip and Thigh have very low importance values, less than 0.5, showing they contribute minimally to the predictive model.

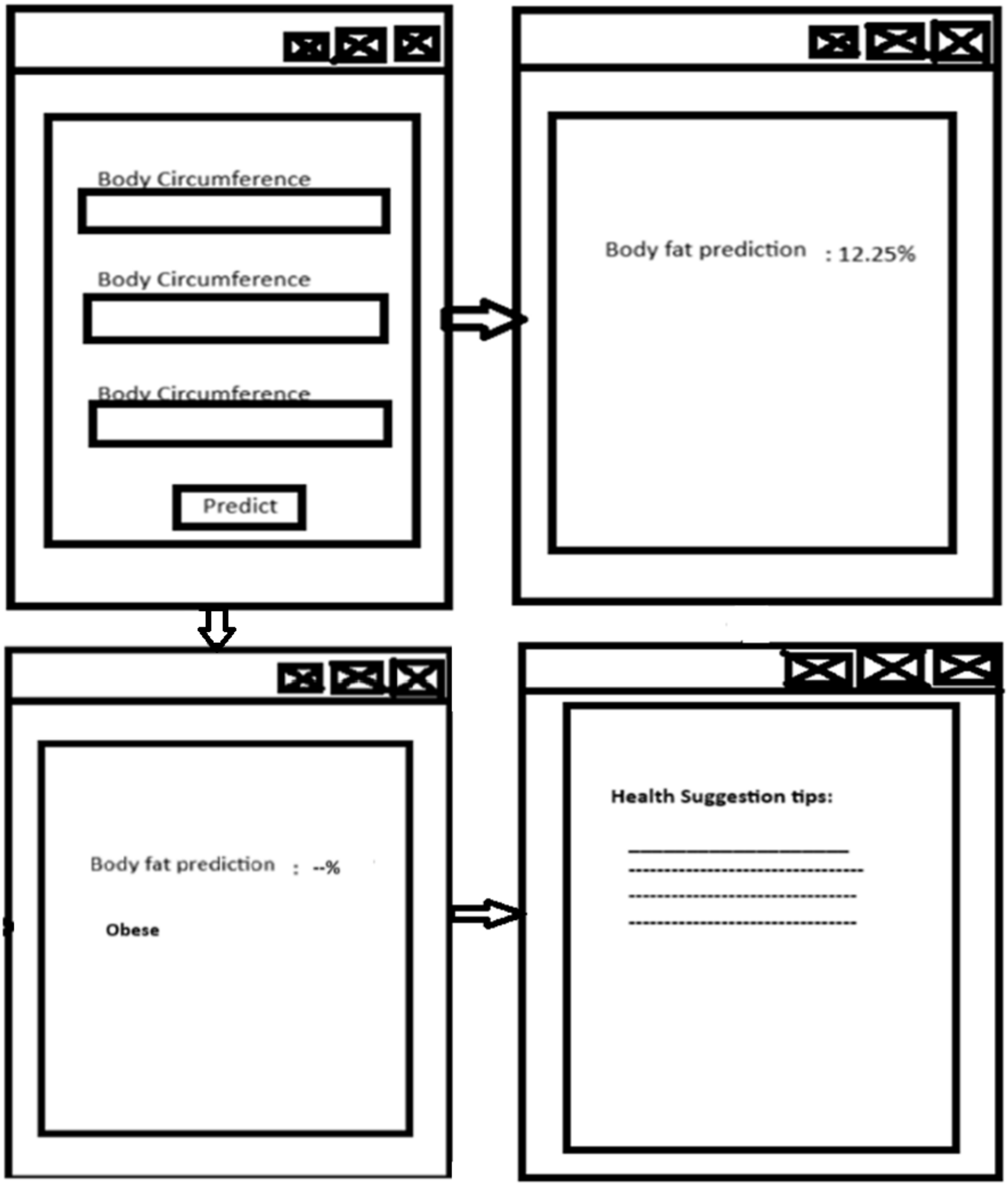
- **Model Training and Evaluation**:

In our project, we trained and evaluated three machine-learning models:

DecisionTreeRegressor, RandomForestRegressor, and SVR (Support Vector Regressor). We chose these models because they are well-suited for regression tasks, offer flexibility in handling different types of data, and allow us to fine-tune parameters to enhance prediction accuracy. Each model brings unique strengths to the project, helping us identify the best fit for predicting body fat percentage. Hyperparameter tuning will be implemented to enhance the model's performance.

- **Web Application Prototype**:

A user-friendly interface for doctors and patients to input various anthropometric measurements, view body fat percentage, and determine an individual as obese, including sections for treating obesity. This app is hosted on the internet to obtain real-time predictions. Figure (d), represents a low-fidelity prototype the prototype contains input for body circumferences, and the second predicts the body fat percentage in individuals. It also provides healthy suggestions to individuals.



Fig(d): Low Fidelity Prototype.

Section 2: Progress to date.

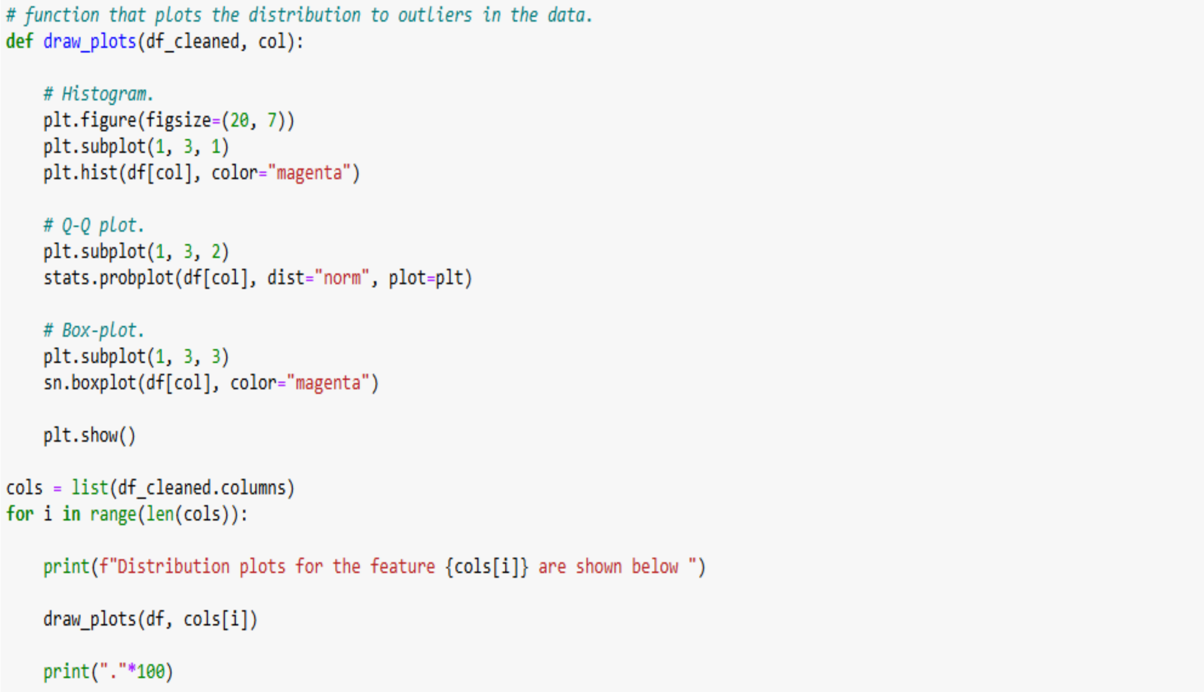
1.) Data Cleaning:

- **Handling Missing Values**: No missing values were found in the dataset, ensuring its completeness and integrity. The dropna() method was applied to verify the absence of missing data, allowing for consistent analysis without the requirement of further imputation or data removal, thus preserving the dataset's original structure and quality.

- **Feature Selection**: To improve model performance, use feature selection techniques. The mutual information gain technique, and ExtraTreesRegressor to identify and top 5 features. Identified the most important features and retained only those. The chest, weight, density, abdomen, and hip were considered as top 5 features. These methods reduce dimensionality, enhancing model interpretability and efficiency.

- **Identifying outliers**: Utilized z-score to identify outliers in the dataset. Boxplots, histograms, and scatter plots were created to visualize the outliers for each feature. After Removing the outliers the size of the dataset was reduced to 242 entries.

- **Distribution plots**: Created distribution plots to identify relationships between the body fat percentage and other anthropometric measurements. There was no relationship recognized between body fat and Density. The body fat and chest circumference represent the gaussian distribution. The abdomen circumference and body fat percentage are gaussian. These can be considered as important features of the dataset.



```
# function that plots the distribution to outliers in the data.
def draw_plots(df_cleaned, col):

    # Histogram.
    plt.figure(figsize=(20, 7))
    plt.subplot(1, 3, 1)
    plt.hist(df[col], color="magenta")

    # Q-Q plot.
    plt.subplot(1, 3, 2)
    stats.probplot(df[col], dist="norm", plot=plt)

    # Box-plot.
    plt.subplot(1, 3, 3)
    sn.boxplot(df[col], color="magenta")

    plt.show()

cols = list(df_cleaned.columns)
for i in range(len(cols)):

    print(f"Distribution plots for the feature {cols[i]} are shown below ")

    draw_plots(df, cols[i])

    print("."*100)
```

- **Model Building:**

  In the initial steps, we created the baseline model to evaluate the performance metrics.

  - **Train-Test Split**: The dataset was divided into two distinct subsets: a training set comprising 70% of the data and a test set containing the remaining 30%. This partitioning allowed us to train the model on a substantial amount of data, enabling it to learn patterns and relationships effectively.

- **Mean Absolute Error (MAE)**: The Mean Absolute Error (MAE) of 6.75 in comparison to the average body fat percentage of 18.96, it appears that the MAE is relatively high. Lower the mae better the performance of the model
- **Root Mean Square Error (RMSE)**: The Root Mean Squared Error (RMSE) of 8.36 in comparison to the average body fat percentage of 18.96, it appears that the RMSE is less than 50%. Lowering the RMSE better the performance of the model.

In the next phase, hyper-parameter tunning will be carried out to enhance the performance of the model. And will be implemented on the different machine learning algorithms.

Section 3: Ethical, Legal, Professional, and Social Considerations

- **Ethics Approval:** Ethics approval is not required for this project as there are no human participants involved. The project primarily uses non-invasive physical measurements, such as body circumference and weight, to estimate body fat percentage. Since the data collected is not related to any medical treatments or interventions, and the application does not involve human participants directly, ethics approval is not necessary. However, ethical considerations remain important, particularly concerning how user data is handled. If the project expands in the future to include more personal or sensitive data, or if it becomes more health-focused, then ethics approval may be required.

- **Ethical Issues:** The main ethical concern is privacy. Even though the data (like chest and hip circumference) are not overly sensitive, they can be linked to personal health information. It is important to collect only the data needed for body fat estimation and ensure it is stored securely. Users must also be informed about how their data will be used. If the application starts giving more personalized health advice, informed consent from users becomes even more important. Another ethical concern is to avoid harm. If the body fat estimates are inaccurate, they could mislead users about their health. Clear warnings should be included in the app to remind users that it is not a substitute for professional medical advice.

- **Legal Issues:** From a legal perspective, the application must follow laws like the General Data Protection Regulation (GDPR) if it is used in the UK. This means ensuring user data is handled transparently, stored securely, and only kept for as long as necessary. A privacy policy must be provided, explaining how user data is collected, used, and protected. Intellectual property rights are also a consideration, ensuring that the app does not infringe on others' copyrights or patents, especially regarding the code and design.

- **Professional Issues:** Professionally, the app must follow industry standards for quality and security. The app should be thoroughly tested to ensure the body fat estimates are as accurate as possible. Accountability in development is important, especially since health-related apps can affect users’ decisions. It’s also important to protect user data by following best practices for security. As the project grows, staying updated on health-tech standards will be essential to ensure the app remains reliable and safe for users.

- **Social Issues:** Socially, the application could have an impact on privacy. Users may be concerned about their health data being collected, so it's crucial to build trust by being transparent about how their data will be used. The app must also be careful not to discriminate or provide inaccurate results for certain groups. Body fat estimation can differ between different genders or ethnicities, so ensuring the model is accurate for all groups is important.

Section 4: Project Plan

- **Remaining Tasks and Timelines**

To ensure the successful completion of the project within the allocated 600 hours, the following tasks are outlined with specific targets and estimated hours for each task. The Initial 280 hours of the project is completed successfully.

1. **Literature Review & DPP Submission (5th October to 19th October 2024, Estimated Hours: 120)**

- Research relevant literature on machine learning models and obesity.
- Summarize findings to inform model design and validation.
- Identify key methodologies and existing applications.
- Review recent studies and trends in obesity and health awareness.

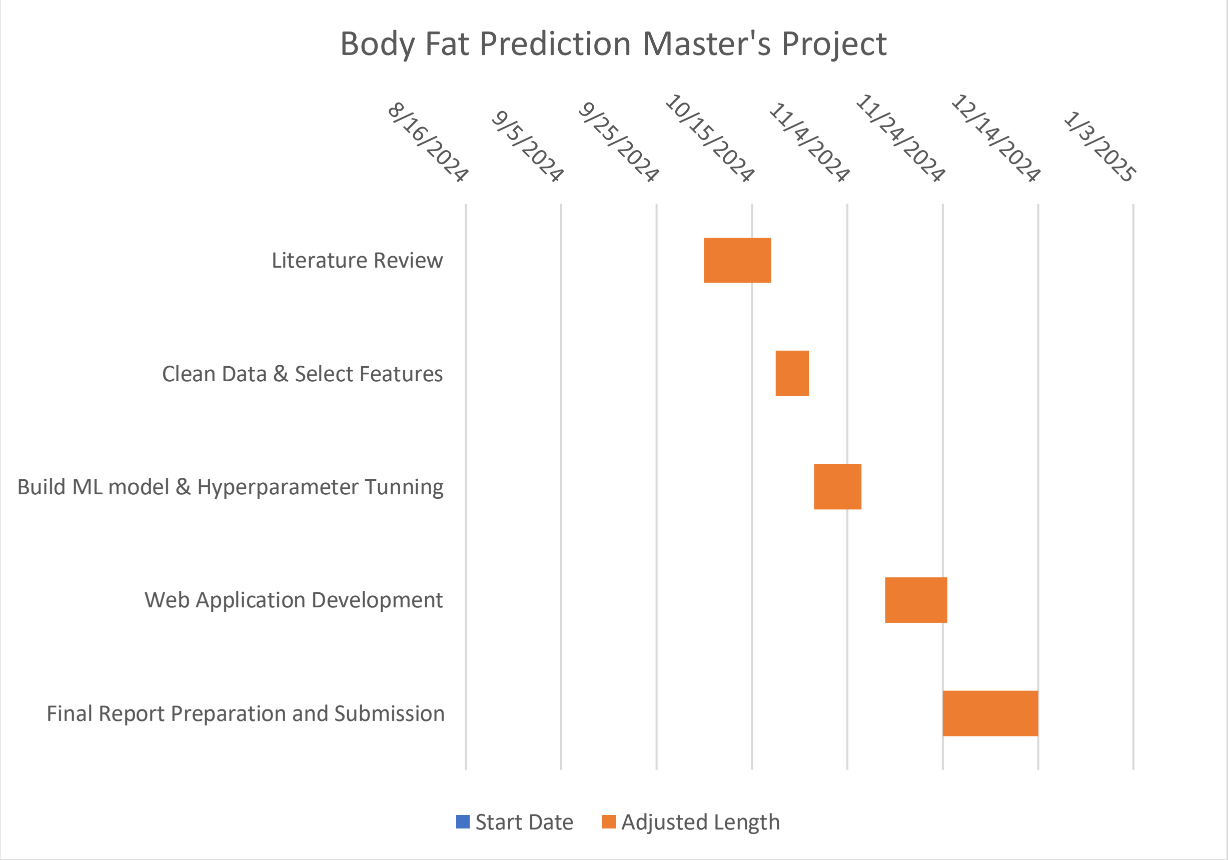
2. **Clean Data & Select the Features (20th October to 27th October 2024, Estimated Hours: 80)**

- Utilize the cleaning function to clean the dataset.
- Implement the Feature Selection algorithm using mutual information gain and ExtraTreesRegressor to select features of high importance.
- Building Baseline model determining MAE, RMSE, etc.

3. **Build ML model & Hyperparameter Tunning (28th October to 11th November 2024, Estimated Hours: 80)**

- Building DecisionTreeRegression, RandomForestRegression, and SVR with hyperparameter tuning.
- Conduct hyperparameter tuning to optimize model performance.
- Perform additional testing with cross-validation to ensure the model's accuracy.
- Evaluate performance using metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

4. **Web Application Development (12th November to 24th November 2024, Estimated Hours: 100)**
   - ○ Convert design mockups into code using HTML, CSS, and JavaScript.
   - ○ Implement responsive design for compatibility across devices (desktops, tablets, mobile).
   - ○ Conduct functional testing to ensure all features work as intended.
   - ○ Implement changes and conduct usability testing to confirm effectiveness.
   - ○ Perform usability testing to gather feedback from users.
   - ○ Address bugs and issues identified during testing.
5. **Final Report Preparation and Submission (Target: 4th January 2025, Estimated Hours: 120)**
   - ○ Compile a final report summarizing project findings, methodologies, and conclusions.
   - ○ Prepare visual aids and materials for the final presentation.
   - ○ Rehearse the presentation to ensure clarity and coherence during the demonstration of the application.



Fig(f): Gantt chart representing project timelines.

Section 5: Level of the Project

This project addresses the crucial problem of obesity, a growing health challenge in the UK and globally, by developing a predictive model that estimates body fat percentage based on non-invasive physical anthropometric measurements. This topic aligns closely with my career aspirations in data science and health informatics, as it combines advanced machine-learning techniques with practical applications in public health. By developing this model, I aim to contribute to the growing need for preventative health solutions, which can assist individuals and healthcare professionals assess obesity-related health risks earlier and more effectively. This project's complexity and real-world application make it a valid MSc-level Project, requiring rigorous testing, and building web applications. This project has a lot of potential for future growth. By adding health suggestions and personalized diet plans, it could become a more complete tool that not only helps users assess their obesity risk but also guides them on how to improve their health. This would align perfectly with the project's goal of preventative care, offering something valuable for individuals and healthcare professionals. Providing personalized recommendations based on the model's predictions would take a more holistic approach to managing obesity. In the future, there's an opportunity to expand the project even further, offering tailored solutions that can help address obesity-related health risks in a more impactful way. The artifact at the core of this project is a machine-learning model that estimates body fat percentage using measurements like chest, abdomen, and hip circumferences. This model is designed to act as an early warning system, helping to predict obesity risks and offering an accessible, non-invasive alternative to traditional body fat estimation methods. The model will be implemented within a user-friendly web application, enabling both patients and healthcare providers to input measurements and receive predictions quickly. This solution improves on existing approaches by offering a more affordable and user-accessible tool, especially for environments where advanced body composition analysis equipment is unavailable. From a technical perspective, the project demonstrates advanced skills by integrating feature engineering, rigorous statistical testing, and model evaluation techniques. I have employed mutual information and ExtraTrees regressor methods for feature selection, ensuring only the most relevant predictors are used to enhance the model's accuracy and reduce potential multicollinearity issues. Additionally, calculating the Variance Inflation Factor (VIF) for selected features ensures the model's stability by reducing redundancy, which is essential for accurate predictions. Through hyperparameter tuning and cross-validation, the model's performance is carefully optimized, balancing predictive accuracy and generalizability.

The depth of the investigation is underscored by an extensive literature review, followed by iterative experimentation and model refinement. I plan to thoroughly test and validate the model's performance using Mean Absolute Error (MAE) , Root Mean Squared Error (RMSE) metrics, etc., to gauge accuracy and reliability.

Building the web application and deploying it on the internet is a time-intensive aspect of this project, demanding considerable coding and development effort. The web application must be carefully designed to ensure a user-friendly experience, intuitive input of measurements, and a clear presentation of the predicted body fat percentage and obesity risk. Additionally, creating a responsive web page is essential, as it allows the application to be accessible and visually consistent across different devices and screen sizes, from mobile phones to desktops.

Section 6: Bibliography:

1. Muhammed Kürşad Uçar, Zeliha Uçar, Fatih Köksal, Nihat Daldal (2021) 'Estimation of body fat percentage using hybrid machine learning algorithms', *Measurement*, vol-167, pp. 1-13.

2. Fedesoriano (2022). *Body Fat Prediction Dataset*. Available at: https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset (Accessed: 01 October 2024).

3. NHS OBESITY (2023): https://www.nhs.uk/conditions/obesity/ (Accessed: 5 November 2024).

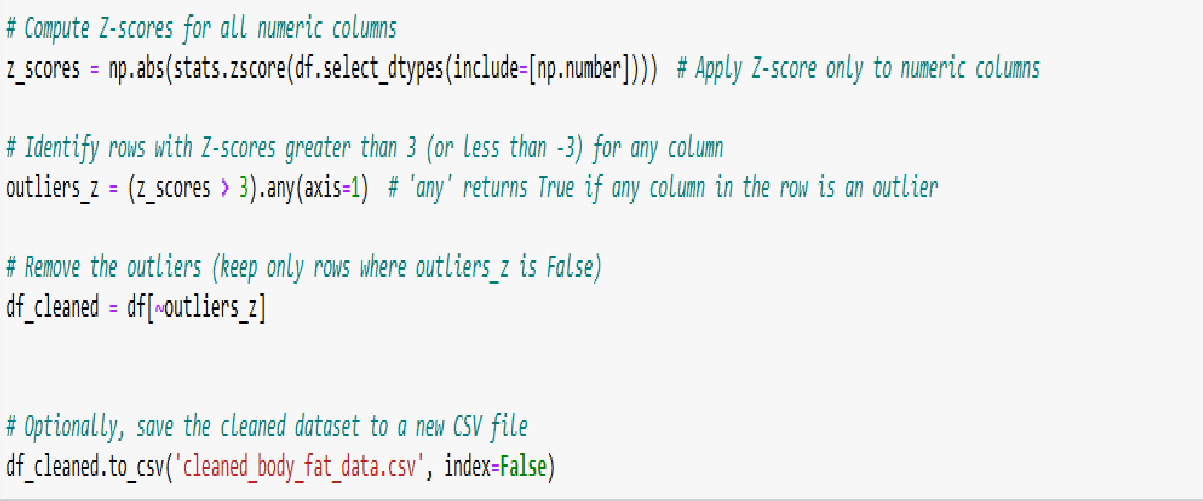
4. Penn Medicine (no date) *Obesity Facts, Definition, and Statistics*. Available at: https://www.pennmedicine.org/for-patients-and-visitors/find-a-program-or-service/bariatric-surgery/who-is-a-candidate/weight-loss-and-obesity-facts (Accessed: 07 November 2024).

Section 7: Appendices

Appendix 1: Data Preprocessing Code

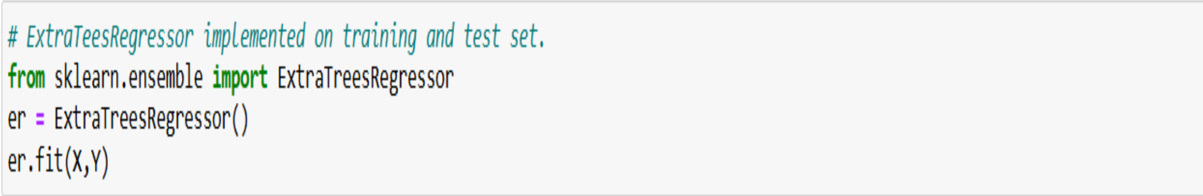
This section contains code snippets and descriptions for data preprocessing, including data cleaning, normalization, and feature selection. Key processes are highlighted:

- **Outlier Detection and Removal**: The code uses scipy.stats for identifying and handling outliers, with visualizations created using matplotlib.pyplot, and seaborn.

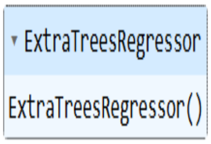


```
# Compute Z-scores for all numeric columns
z_scores = np.abs(stats.zscore(df.select_dtypes(include=[np.number])))  # Apply Z-score only to numeric columns

# Identify rows with Z-scores greater than 3 (or less than -3) for any column
outliers_z = (z_scores > 3).any(axis=1)  # 'any' returns True if any column in the row is an outlier

# Remove the outliers (keep only rows where outliers_z is False)
df_cleaned = df[~outliers_z]


# Optionally, save the cleaned dataset to a new CSV file
df_cleaned.to_csv('cleaned_body_fat_data.csv', index=False)
```

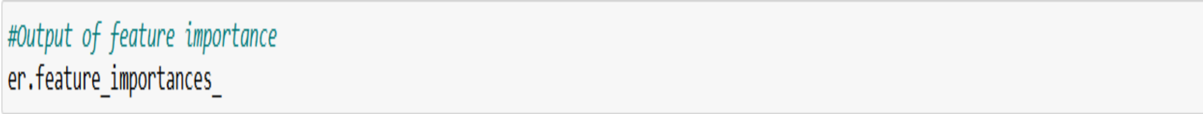
- **Feature Selection**: Using mutual information and ExtraTrees Regressor, the code ranks and selects the top features impacting body fat prediction.



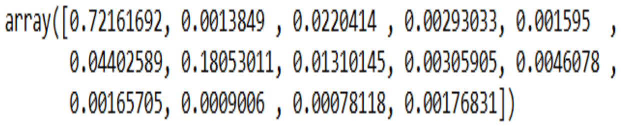
```
# ExtraTreesRegressor implemented on training and test set.
from sklearn.ensemble import ExtraTreesRegressor
er = ExtraTreesRegressor()
er.fit(X,Y)
```



```
▾ ExtraTreesRegressor
ExtraTreesRegressor()
```



```
#Output of feature importance
er.feature_importances_
```



```
array([0.72161692, 0.0013849 , 0.0220414 , 0.00293033, 0.001595  ,
       0.04402589, 0.18053011, 0.01310145, 0.00305905, 0.0046078 ,
       0.00165705, 0.0009006 , 0.00078118, 0.00176831])
```

Appendix 2: Machine Learning Model Results

This appendix provides a summary of the machine learning models tested, along with their performance metrics:

- **Results Summary Table**:

Model	**MAE**	**RMSE**	**R² Score**
Decision Tree	1.16	1.49	0.96
Random Forest	0.61	0.71	0.99
Support Vector Regressor	4.84	6.05	0.28

These results show that the Random Forest model achieved the highest accuracy and reliability in predicting body fat percentage.