1. **Briefly describe your methodology for analyzing the data.**

First and foremost, I read the train and test datasets in order to have a comprehensive understanding of them. Is there any missing or duplicate values in the dataset? I eliminated them because there were very few. The most missing values were in the weight columns, and the rest of the columns had less missing values than the weight columns, but all of the columns had missing values. For a more in-depth understanding, I demonstrated the data into various graphs and attempted to comprehend the feature conditions and their relationships with one another. Then, to improve prediction, I invented a new feature called "BMI". It accurately depicts a person's body composition, which directly determines body measurements. After that, I discovered whether outliers are important or not, and I concluded that outliers are true, valid aberrations in human body shape. It is better to utilize Random Forests for a realistic and inclusive model because they are robust to outliers and the objective is to predict for a variety of individuals. encoding gender and other category characteristics. Next, I used StandardScaler to standardize numerical features like weight and height. I used GridSearchCV to optimize hyperparameters and trained regression models (such as Random Forest) to predict body measures. Lastly, I tested predictions on fresh input data and assessed model performance using MAE, RMSE, and R2.

2a. Which ML algorithm did you pick for making predictions? Why?

I'll utilize the Random Forest Regressor to make predictions since it can handle non-linear relationships and uses ensemble approaches, in which the training output of several models is derived from the average of those models. It has therefore become less susceptible to overfitting, more robust, and its variance has decreased. This model offers baseline performance and robustness to outliers, requires little adjustment of hyperparameters, and can handle missing values. It frequently handles datasets with imbalanced classification and feature importance assessment, making it an extremely useful model.

2b. Which ML algorithm would be a poor choice for this dataset? Why?

Because KNN struggles to compute the distance (Euclidean, Manhattan, Minkowski, etc.) between the test and training samples and produces slow prediction times for production usage, I thought it would be a bad choice for this dataset, even if it were enormous. This dataset contains a variety of features, including height, weight, age, and gender. Distance measures used in KNN lose their significance as the number of attributes increases. Additionally, KNN will be biased if numerical features such as weight and height are not appropriately normalized. Furthermore, this dataset has outliers in the features of height, weight, bust, waist, and hip, making it extremely sensitive to misleading and outlier data. Complex class faces and class imbalance were beyond KNN's capabilities.

3a. Which evaluation metric did you pick? Why?

MAE is used to interpret the average absolute difference between expected and actual values, which is why I chose it. Furthermore, it calculates the amount of incorrect predictions I made, is simple to interpret, is insensitive to outliers, and offers a robust error measurement. It is also computationally efficient and offers a more understandable estimate of average error. Additionally, when the data is normally distributed, RMSE computes the square root of MSE and indicates the average number of incorrect predictions. The amount of variance captured in the target variable was measured by R2. The greater the value, the better the model fits, the baseline performance, and the normalized measure.

3b. What evaluation metric would you use if this was a classification problem to predict whether a person was male or female?

I thought the primary focus would be accuracy, since it explicitly states the percentage of correctly identified samples among all samples. Accuracy can be a useful indicator with comparable outcomes for false positives and false negatives.

3c. How would your answer change if the classes were very imbalanced?

When accuracy is insufficient due to imbalance in the dataset, I would like to take into account precision, recall, F1-score, and the AOC-ROC curve, as these would provide me with more useful information. They take into account the model's performance on the minority class. For example: if 90% of the data are male, the model will consistently predict "Male" with greater accuracy, even if it fails to identify every instance of a female.

4. The body measurement prediction problem is a regression problem. In a classification problem which would worry you more: false positives or false negatives? Why?

It would be more worrying to have false positives in a classification scenario of the body measurement prediction problem. It is possible to cause discomfort, discontent, and returns if someone is predicted to fit a smaller size than they actually do. These are all serious problems in the fashion retail industry. A looser fit could result from false negatives, which is less dangerous in contrast.

5. How would you deal with missing data for numerical and categorical variables?

Numerical:

Delete— Remove the rows if the missing values are less than 5% and the data is randomly missing. **Mean**— If the distribution of the missing dataset is symmetrical and normal. Additionally, there are no outliers and feature values are numerical. **Median**—When data is skewed or contains outliers, utilize median imputation. **KNN Imputation**—The values of the k-nearest neighbors in the feature space are used to impute missing values. **Interpolation**— Use either linear or polynomial interpolation for time series data. **Random**— Use any random value that isn't part of the current feature to fill in the missing data. for numerical use (-1,0, or similar to 99.99) and for category fill with the "Missing" key word. **Arbitrary**—When data is not missing at random, arbitrary can be used. **End of distribution**— Add the final values from the dataset distribution. **Iterative Imputer (MICE)** — uses the associations between features to construct a sequence of regression models to forecast missing values. It works best when the data is randomly missing and the features are associated.

Categorical:

Delete—Rows should be deleted when the percentage of missing data is less than 5% and when the data is randomly missing. **Mode**— Use the most common category to substitute missing values. **Random Imputation**—This method works best for linear models. Choose a random integer from the dataset and enter "Na" values. **Create a New Category**—Give the missing values a new, unique category. For instance, under a 'Color' column, you may substitute a new category such as 'Missing' or 'Unknown' for all NaN or empty strings.

6. Explain regularization to a layperson. What is it and why would you want to use it?

Assume for a moment that I am preparing a dish for my family by following the recommendations of each member. If my father says "Use garlic instead," my brother says "Too spicy," and my mother adds "Add more salt," I will do as they say. At last, the dish becomes a mess. I should pay more attention to the important contents than the noise.

Technically:

It's a machine learning technique that keeps a model from learning too much about the training data, such as noise and unimportant patterns, which might cause it to perform poorly on new, unseen data.

Why use it:

Please, Shehab, don't try to use every item in your dish. functions similarly to a "Discipline Rule" for the model. Just the most important ingredients, not too much.

Technically:

Make the model simple and accurate, enhance generalization, and eliminate irrelevant features to avoid overfitting.

7. What’s the difference between L1 and L2 regularization methods?

In order to avoid overfitting, both L1 and L2 regularization include a penalty term in the loss function that was used to train the model.

1. The sum of the absolute values of the coefficients is added to the loss in L1, and the sum of squares is utilized in L2.
2. Some weights are made exactly to zero by L1, other weights are gently shrunk by L2.
3. While L2 is not used, L1 is used for feature selection.
4. L2 is quicker computationally than L1.
5. L2 employs a circular/elliptical-shaped restriction region, while L1 utilizes a diamond-shaped one.
6. L2 is simpler to compute and L1 is more challenging to optimize.

8. What is better: 50 small decision trees or one large decision tree? Why?
Ans:

50 small decision trees like Random Forests or Boosted Trees are preferable. In comparison to a single huge tree, which can overfit, provide poor generalization, and be difficult to read and comprehend. On the other hand, they better capture complicated patterns, handle multicollinearity, reduce overfitting, produce stable predictions, and generalize effectively.