Predicting Body Fat Percentage Based on Body Circumference Measurements

IN R

Pei Yii Ng (Heather)

11 November, 2021

BACKGROUND

High body fat percentage leads to develop obesity-related diseases, including heart disease, high blood pressure, stroke, and type 2 diabetes.

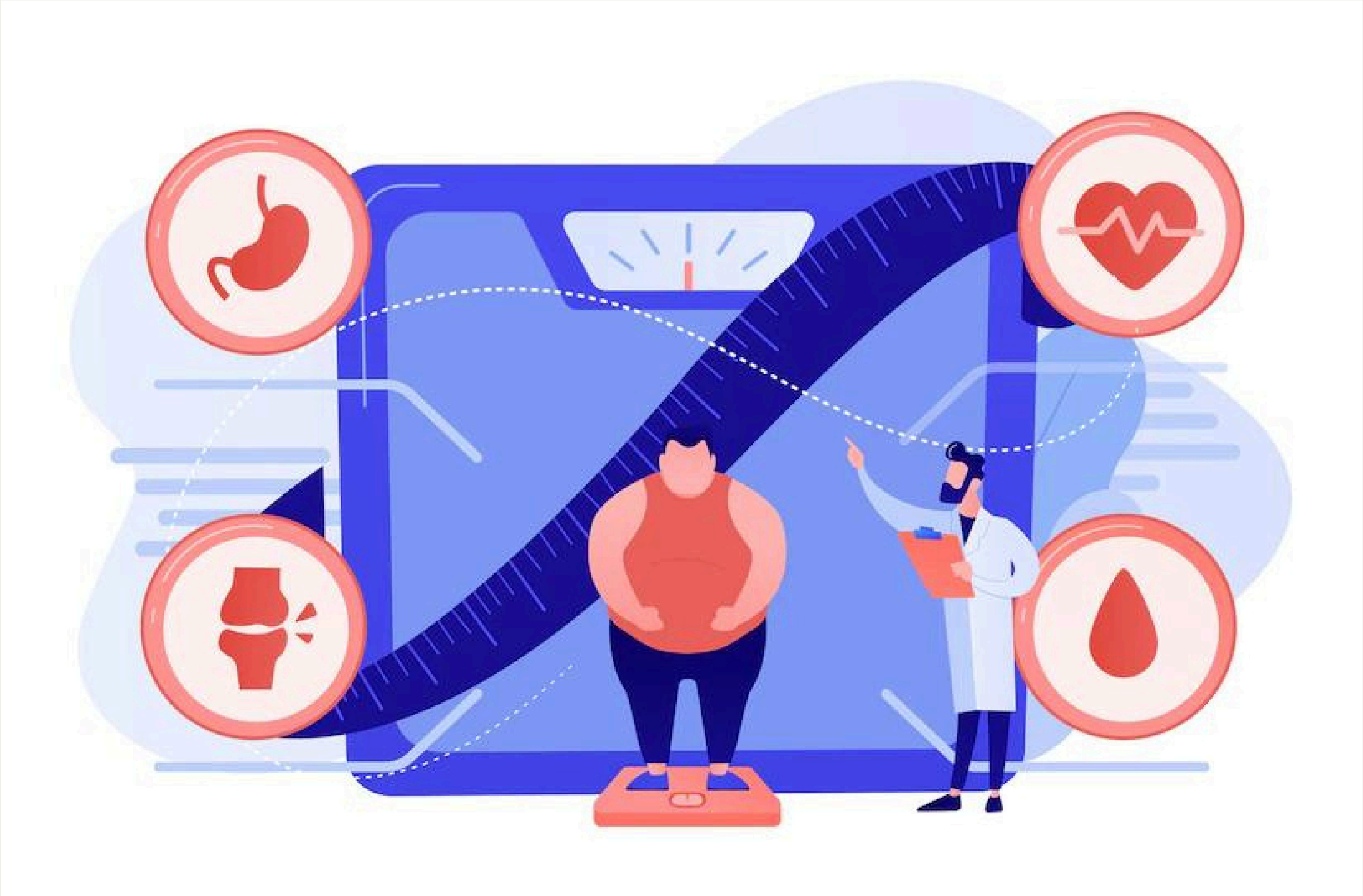
ANALYSIS OBJECTIVES

- Build predictive models to predict body fat percentage from body circumference measurements
- Allow easier estimation of body fat percentage
- Find correlations between variables

DATA SOURCE

Kaggle's 'Body Fat Prediction dataset' by Dr. A. Garth Fisher, provides body fat estimates and various body circumference measurements.

Kaggle Source



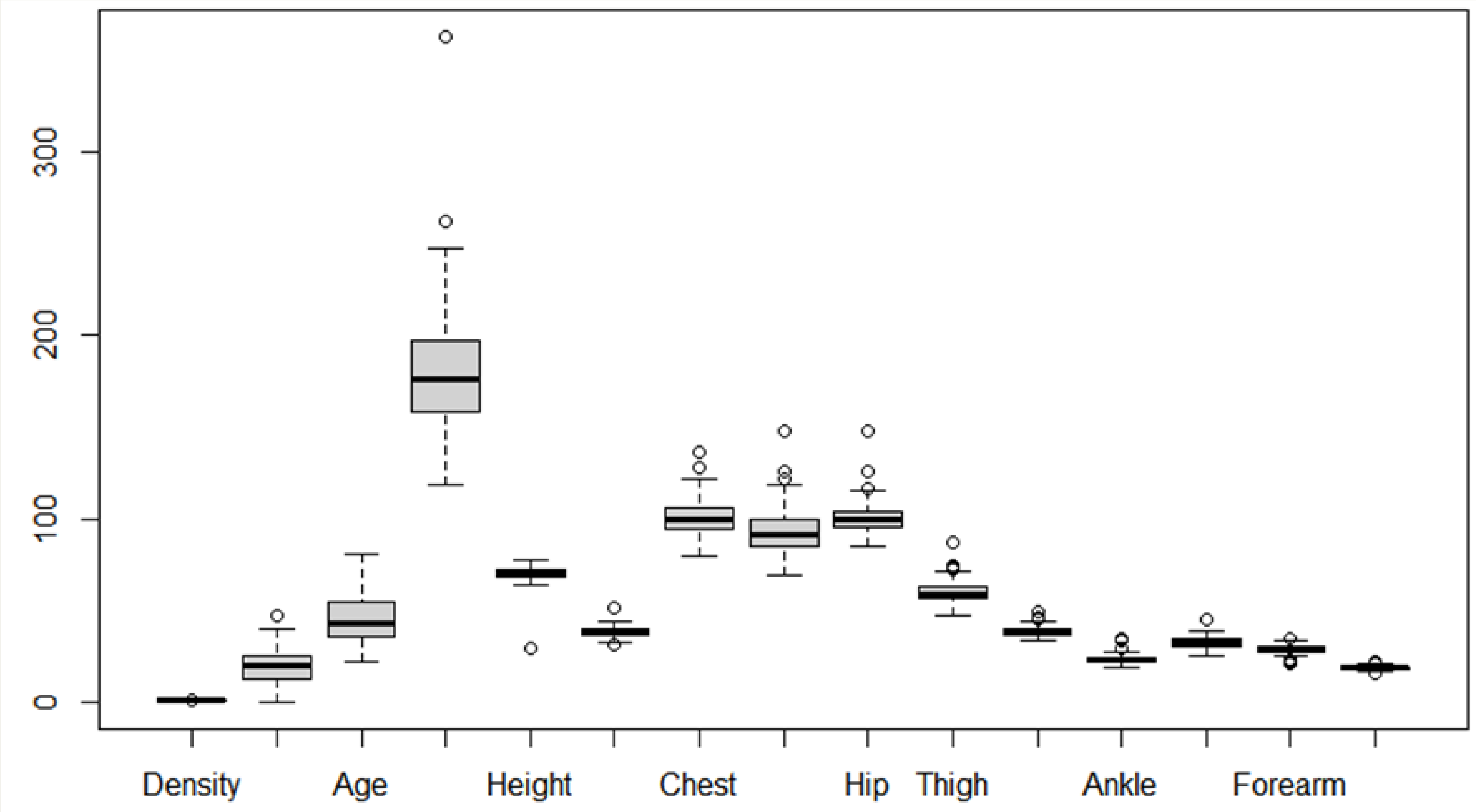
Data content:

1. 252 records of men with 15 attributes each:
   a. Density, Body Fat percentage, Age, Weight, Height
   b. Circumferences: Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist
2. Circumferences in cm, Height in inches

EXPLORATORY DATA ANALYSIS

Issues identified:

- Missing data - zero values in body fat column
- Inconsistent units (height in inches, the rest in cm)
- Outliers detected through box plots



Box plots for each variables

DATA PREPROCESSING

Data Cleaning

impute inaccurate data

- Replace the outliers with NaN
- Replace NaN or Zero to mean value

Data Reduction

Feature selection to select most significant label

- Utilised Boruta library's built in function
- All attributes are considered important, no attributes removed.

Data Partitioning

Implement train-test split

- Use caTools library
- Split dataset to 80% for training, 20% for test

MODELLING TECHNIQUES

Build regression models for predicting body fat percentage as it is a continuous value

- Model 1: Multiple Linear Regression
  - Basic linear model, simple and interpretable
- Model 2: Support Vector Regression (SVR)
  - Handles non-linear problems, robust to outliers
- Model 3: Random Forest Regression
  - Ensemble of decision trees, robust to non-linearity

MODEL 1 - MULTIPLE LINEAR REGRESSION

Principles: Fits a linear equation of predictors to the response variable

Advantages: Simple, easy-to-interpret coefficients

Model performance:

- R-squared: 0.972 (97.2% variance explained)
- Residual Standard Error: 1.383

```
Call:
lm(formula = BodyFat ~ ., data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3211 -0.4259 -0.0909  0.3053 14.6743 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.486e+02  1.387e+01  32.344   <2e-16 ***
Density     -4.156e+02  9.437e+00 -44.036   <2e-16 ***
Age          1.950e-02  1.181e-02   1.652    0.100    
Weight       6.563e-03  2.474e-02   0.265    0.791    
Height       4.000e-02  7.211e-02   0.555    0.580    
Neck         5.308e-03  8.792e-02   0.060    0.952    
Chest        4.180e-02  3.724e-02   1.123    0.263    
Abdomen     -9.325e-03  4.007e-02  -0.233    0.816    
Hip          2.630e-02  4.741e-02   0.555    0.580    
Thigh        3.917e-02  4.663e-02   0.840    0.402    
Knee        -1.582e-02  9.370e-02  -0.169    0.866    
Ankle       -1.550e-01  1.209e-01  -1.282    0.201    
Biceps      -7.523e-02  6.222e-02  -1.209    0.228    
Forearm      2.790e-02  1.099e-01   0.254    0.800    
Wrist        8.954e-02  1.957e-01   0.458    0.648    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.383 on 186 degrees of freedom
Multiple R-squared:  0.9739,    Adjusted R-squared:  0.972 
F-statistic:   496 on 14 and 186 DF,  p-value: < 2.2e-16
```

MODEL 2 - SUPPORT VECTOR REGRESSION

```
> # Model building
> svr_model = svm(BodyFat ~ ., normalized_training_set)
> summary(svr_model)

Call:
svm(formula = BodyFat ~ ., data = normalized_training_set)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.07142857
    epsilon:  0.1


Number of Support Vectors:  94
```

Principles: Maps data to higher dimension, finds optimal hyperplane
Advantages: Effective on non-linear data, handles outliers well
Model performance:

- Mean Absolute Error (MAE): 1.21
- Root Mean Squared Error (RMSE): 1.54
- R-squared: 0.89 (89% variance explained)

MODEL 3 - RANDOM FOREST REGRESSION

```
> rf_model

Call:
 randomForest(formula = BodyFat ~ ., data = training_set)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 4.855137
                    % Var explained: 92.85
```

Principles: Ensemble of decision trees, selects random subsets of features
Advantages: Robust to outliers, automatic handling of non-linear patterns
Model performance:

- Mean of squared residuals: 4.85
- R-squared: 0.9285 (92.85% variance explained)

PERFORMANCE COMPARISON

Model	R-Squared (%)
Multiple Linear Regression	97.2
Support Vector Regression	89.0
Random Forest Regression	92.9

The higher the R-Squared, the more variance it can explain and the more accurate the model is.

CONCLUSION

- Multiple Linear Regression model showed highest accuracy (97.2%)
- All three models exhibited high predictive performance
- Can reliably estimate body fat percentage from circumference measurements
- Linear regression assumptions seem valid for this dataset