Executive Summary

This report details the development of a machine learning model to predict the correctness of weightlifting exercises (`classe`) based on sensor data. The analysis uses the "Weight Lifting Exercises Dataset." After extensive data cleaning to remove variables with missing values and non-predictive metadata, the training data was partitioned into training (75%) and validation (25%) sets. Two models were trained and compared using 5-fold cross-validation: a Decision Tree and a Random Forest. The Random Forest model demonstrated vastly superior performance, achieving over 99% accuracy on the validation set. This high-performing model was selected as the final model, with an estimated out-of-sample error of less than 1%. The final model was then used to predict the `classe` for 20 test cases.

Introduction

The goal of this project is to build a model that can accurately predict *how* a person is performing a weightlifting exercise based on data collected from sensors. The outcome variable, `classe`, indicates the quality of the exercise, with class 'A' being the correct execution and classes 'B' through 'E' representing common mistakes. An accurate predictive model could be used in wearable devices to give users real-time feedback on their form, preventing injury and improving workout effectiveness.

Data Processing and Exploratory Analysis

The analysis was performed on the `pml-training.csv` dataset. The raw data contains 160 variables, many of which are unsuitable for modeling in their original state.

Data Cleaning

The first and most critical step was data cleaning. Upon inspection, a large number of columns were found to contain almost entirely missing values (`NA` or blank entries). These columns provide no useful information for prediction and were removed from both the training and testing datasets. Additionally, several metadata columns—such as subject name, timestamps, and window IDs—were removed as they are not relevant predictors of exercise quality. This cleaning process significantly reduced the number of predictor variables, leaving a smaller, more robust set of features for model training.

Data Partitioning

To properly evaluate model performance and estimate the out-of-sample error, the cleaned training dataset was partitioned into two smaller sets:

- A **training set** (75% of the data) used to build and train the predictive models.
- A **validation set** (25% of the data) held aside to test the models on unseen data.

This split ensures that our evaluation of the models is unbiased and reflects how they would perform on new data.

Modeling Strategy

The primary goal is classification—predicting the correct `classe` (A, B, C, D, or E). To achieve this, two different machine learning models were trained and compared.

1. **Decision Tree (rpart):** A single decision tree was trained as a baseline model. It is relatively simple and easy to interpret but often lacks the predictive power of more complex methods.
2. **Random Forest (rf):** A Random Forest model was trained as the primary candidate model. As an ensemble method that builds hundreds of decision trees, it is known for its high accuracy and robustness against overfitting, making it well-suited for datasets with many predictors.

For both models, a **5-fold cross-validation** strategy was used during training. This technique involves splitting the training data into 5 "folds," training the model on 4 folds, and testing it on the 5th, repeating this process 5 times. This helps to produce a more reliable and generalizable model.

Results and Model Selection

The performance of the two models was compared using their accuracy on the held-out validation set.

- The **Decision Tree** model achieved an accuracy of approximately **73%** on the validation data.
- The **Random Forest** model achieved a significantly higher accuracy of over **99%** on the validation data.

Due to its vastly superior predictive performance, the **Random Forest was selected as the final model**. The accuracy on the validation set provides an estimate of the model's performance on new, unseen data. The estimated out-of-sample error for the final Random Forest model is approximately **0.6%** (calculated as 1 - 0.994).

Conclusion

The Random Forest algorithm proved to be an extremely effective model for predicting exercise quality from sensor data, with an expected accuracy of over 99%. The high accuracy and low estimated out-of-sample error suggest that this model can be reliably used for the intended application.

The final predictions for the 20 cases in the official test set were generated using this model.