Health Tales - A Comparative Study of Predictive Models

Adit Doshi
Yash Kharade
Harshaditya Mallipudi
Prathik Makthala

Introduction

- Exploring how lifestyle choices like smoking and drinking affect health using data-driven models
- Data-driven study aimed at understanding the impact of lifestyle choices, such as drinking and smoking on health
- Objective: To analyze health metrics and provide insights into the consequences of individual habits.

Data Source and Overview

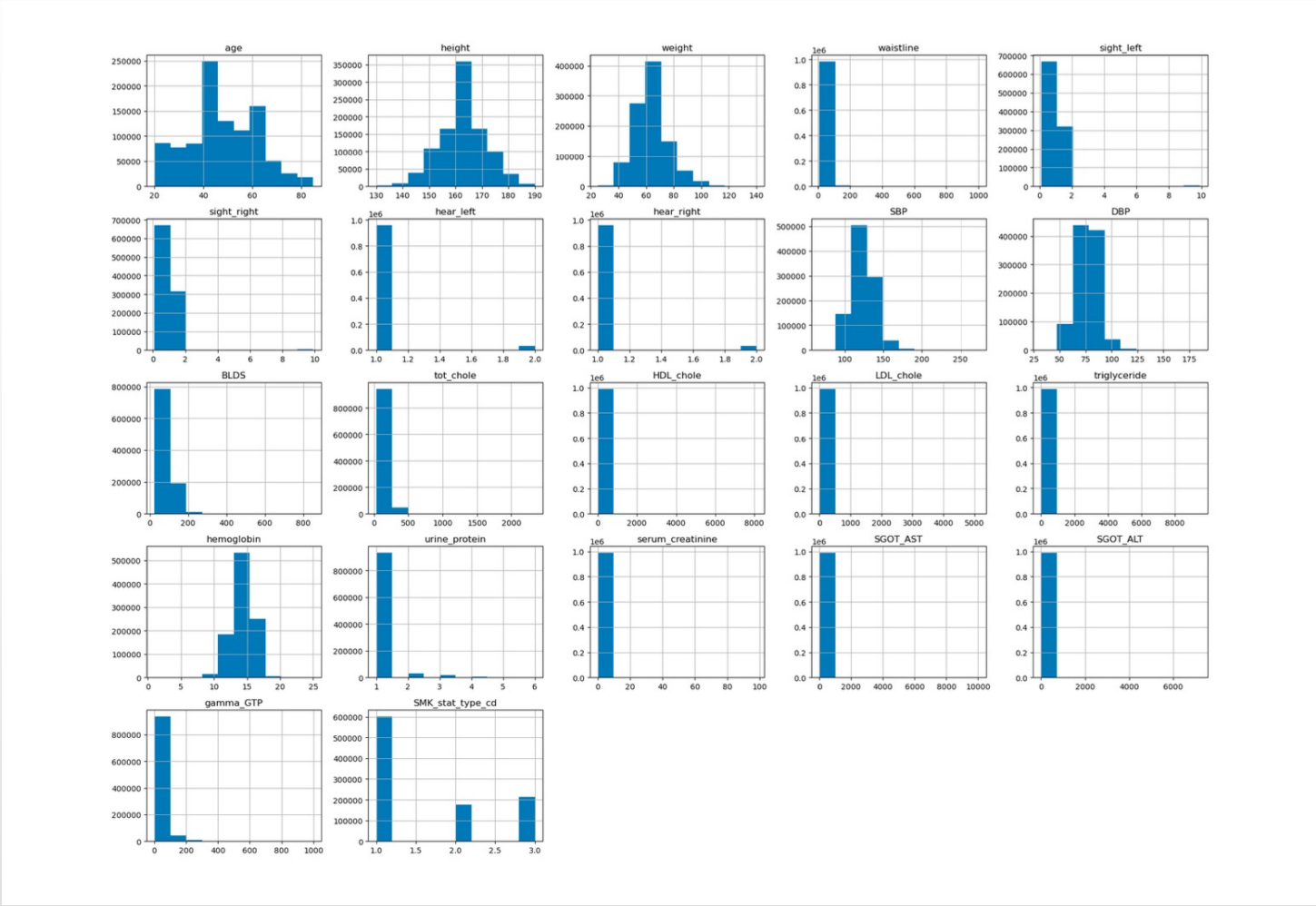
- National Health Insurance Service, Korea.
- Dataset Features: 23 variables including blood pressure, cholesterol, liver function, and lifestyle factors.
- A mixture of qualitative & quantitative variables.
- The purpose of this dataset is to:
  1. Analysis of body signal
  2. Classification of smoker or drinker

Data Cleaning and Preprocessing

- Identified and addressed outliers in waistline, cholesterol levels, etc
- Employed scaling and encoding to prepare data for analysis.
- No missing values were present in the dataset.

Exploratory Data Analysis (EDA)

- Analyzed skewness in data distribution for various health parameters
- Used histograms and correlation matrices to uncover initial insights

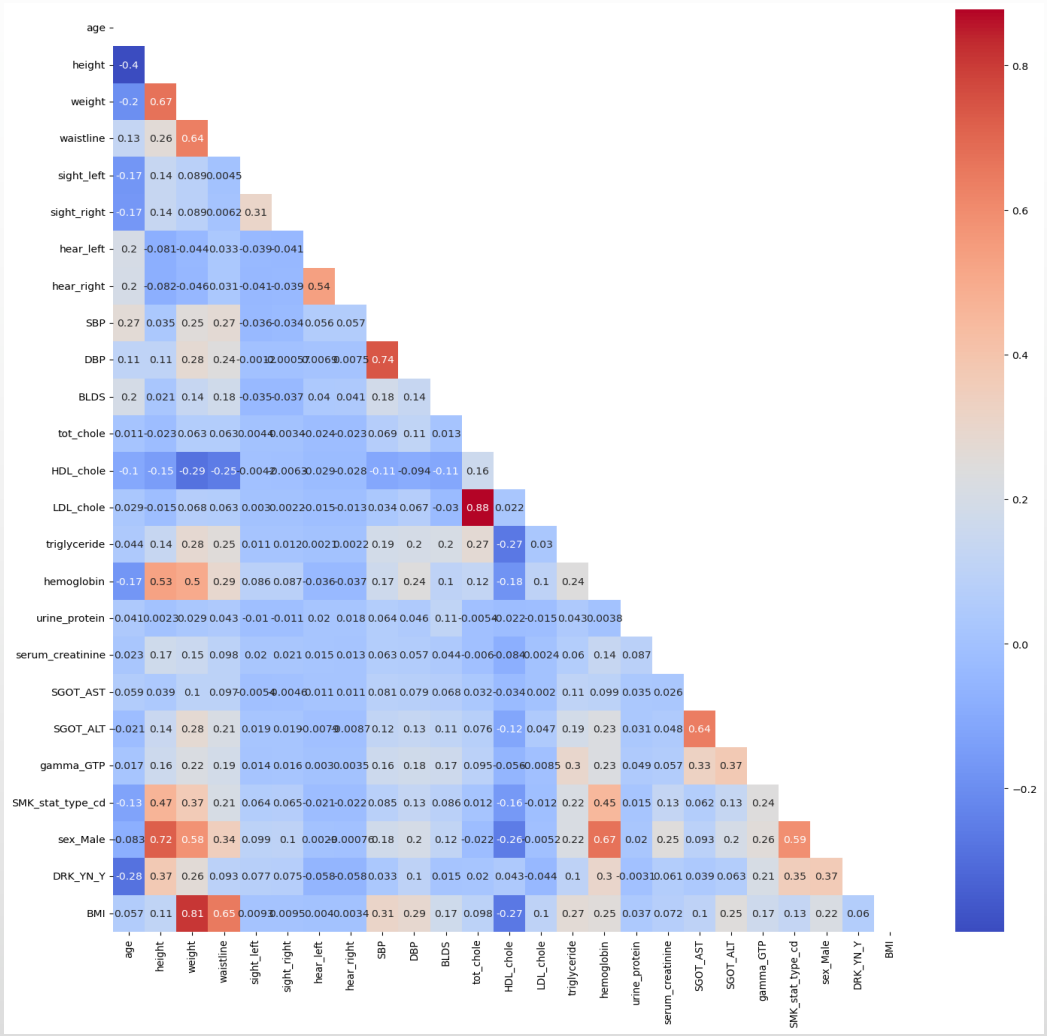


Data Distribution Analysis

- **Symmetric:** Age distribution is even.
- **Right-Skewed:** Higher values for Weight, Blood Pressure (SBP, DBP), Cholesterol (Total, LDL), Triglycerides, Hemoglobin.
- **Strongly Right-Skewed:** Elevated readings in Waistline, Vision and Hearing (Left/Right), Blood Sugar, HDL Cholesterol, Serum Creatinine, Liver Enzymes (SGOT/AST, gamma_GTP).
- **Left-Skewed:** More female participants (indicated by 'sex_Male' skewness).

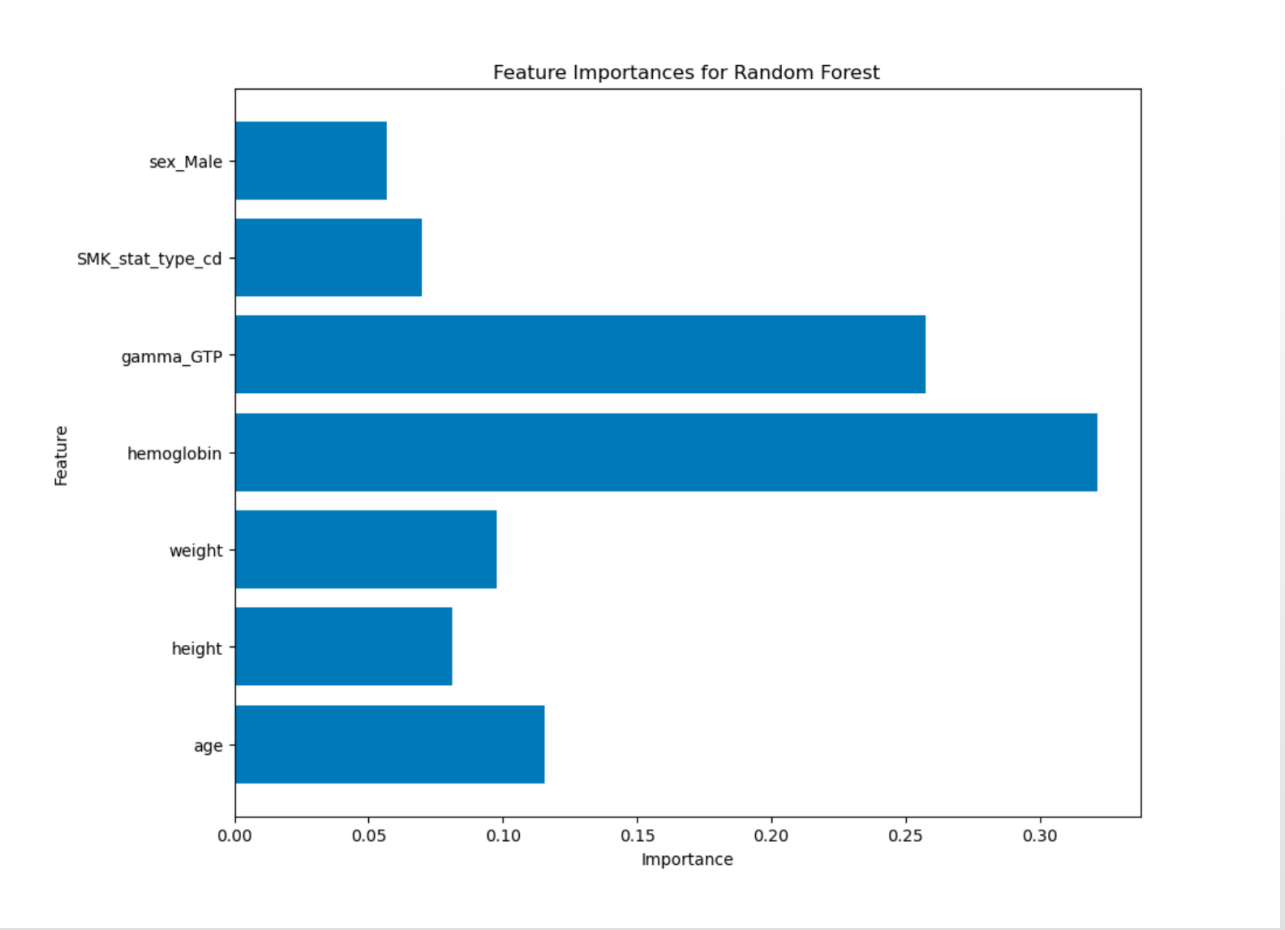
Exploratory Data Analysis (EDA)

- Direct link between weight and waistline, confirming health hypotheses.
- Identified weak correlations to refine predictor variables, eliminating redundancies.



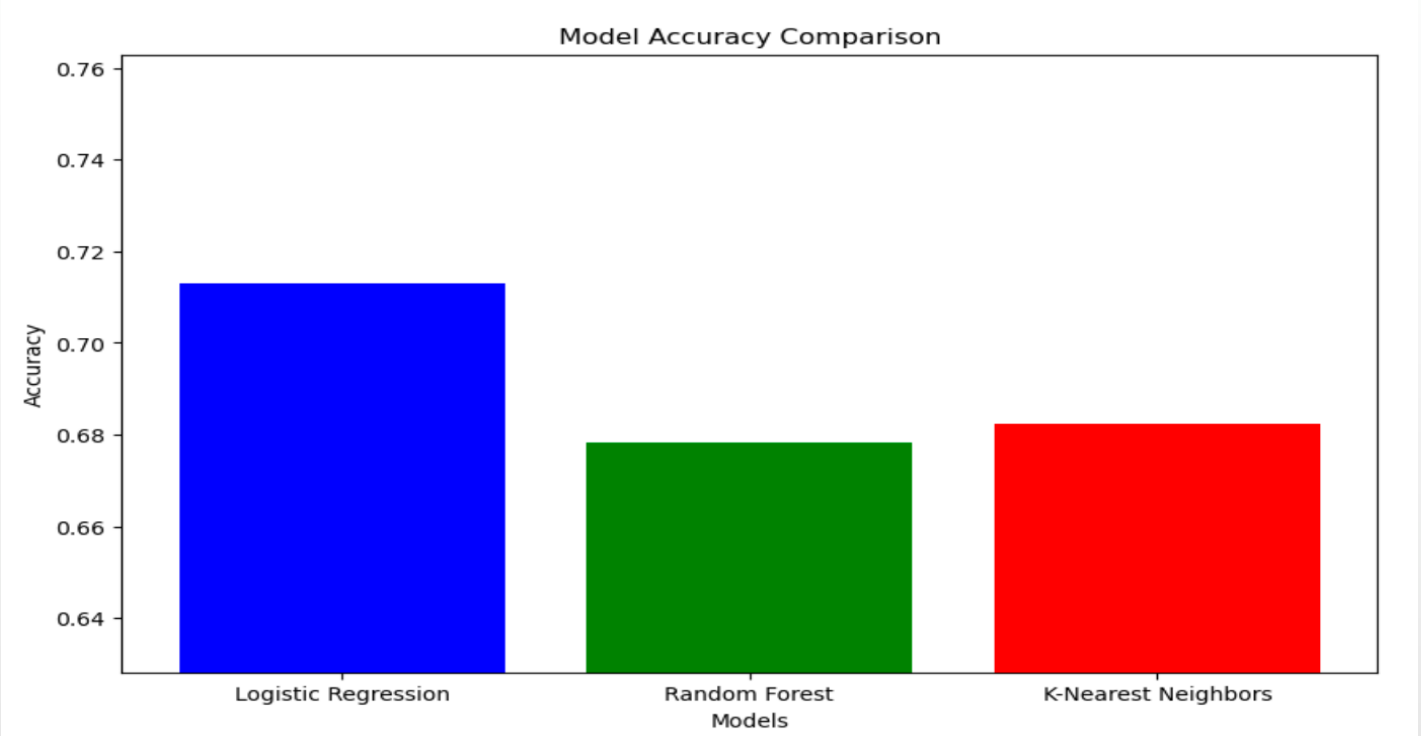
Feature Importance Analysis

- Evaluated the most predictive features in the Random Forest model
- Visualized feature importance for better understanding



Initial Model Training

- Models: Logistic Regression, Random Forest, K-Nearest Neighbors
- Initial Results: Logistic Regression outperformed with an accuracy of 71.29%
- Preliminary Analysis:
  - Logistic Regression: Accuracy – 71.29%
  - Random Forest: Accuracy – 67.86%
  - K-nearest neighbor: Accuracy – 68.23%

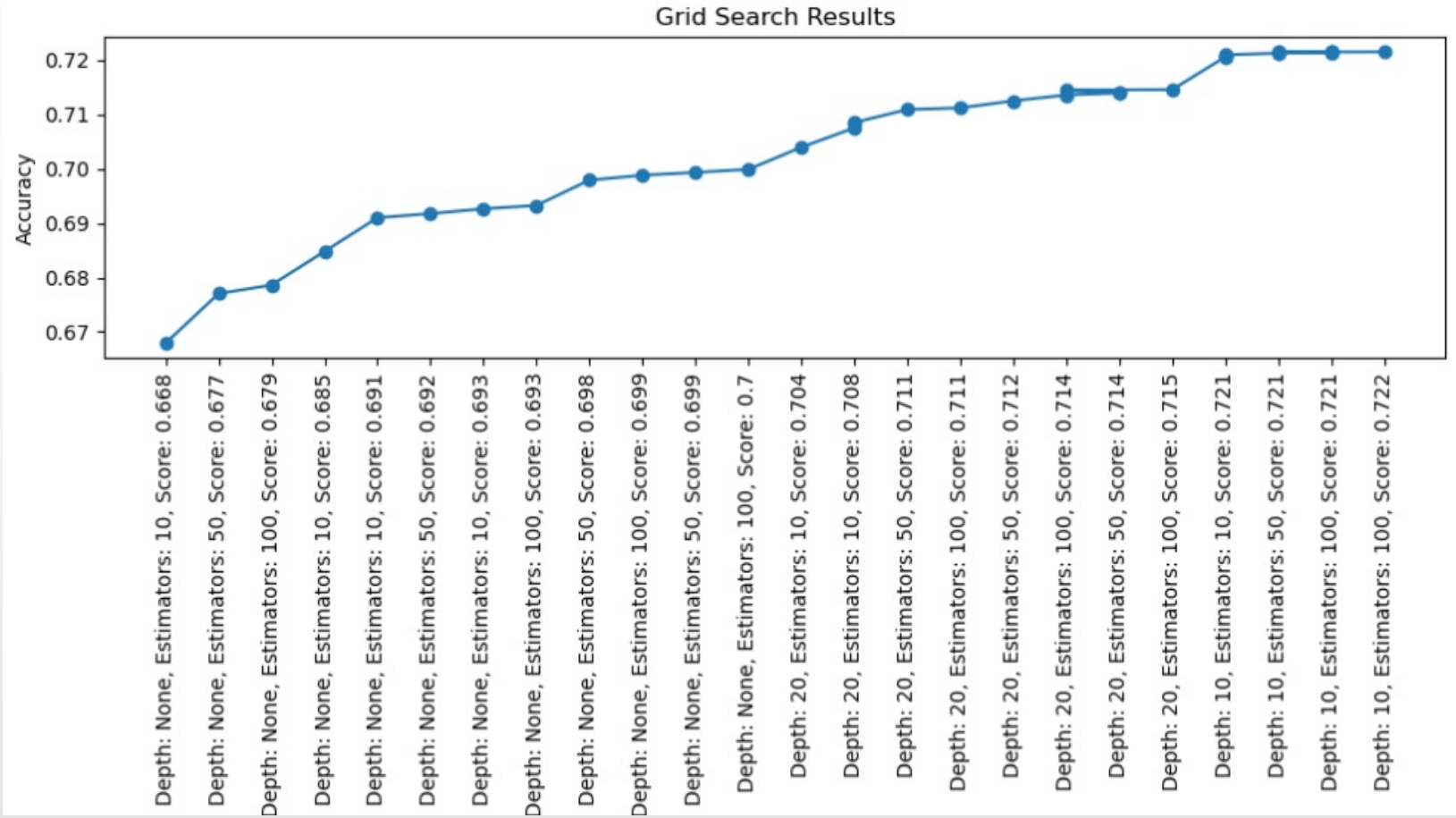


Observations

- The Logistic Regression model has the highest accuracy, followed by K-Nearest Neighbors and Random Forest.
- Precision, recall, and F1-scores are similar for all three models, suggesting that they perform similarly in terms of correctly classifying instances from both classes.
- The ROC scores for all models are also similar, indicating comparable discrimination abilities.

Hyperparameter Tuning and Improved Models

- Applied GridSearchCV for tuning model parameters.
- Achieved enhanced model performance, particularly in Random Forest

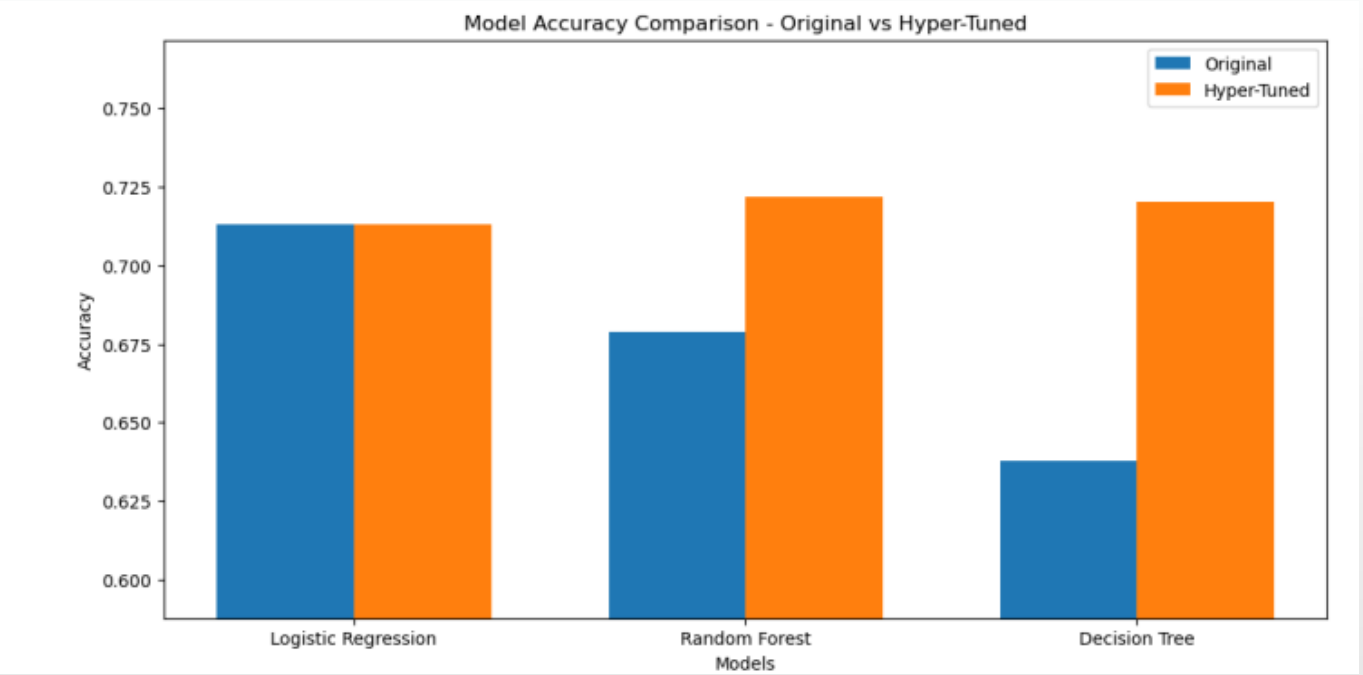


Observations

- The graph shows a positive trend in model accuracy as the number of estimators increases.
- Parameter Impact:
  - Estimators: Accuracy improves significantly as we move from 10 to 100 estimators, demonstrating their strong influence on model performance.
  - Depth: Variations in depth (None, 1, 2) at 100 estimators show negligible differences in accuracy, suggesting depth has a limited effect in this context.
- Depth: None' with 'Estimators: 100' appears to be the most efficient parameter set within the tested range.

Comparative Analysis - Model Performance

- Comparison of model accuracies pre- and post-tuning.
- Decision Tree model showed significant enhancement.
- Random Forest achieved the highest overall accuracy.



Conclusion

- **Weight-Waistline Correlation**: Confirmed the direct association as per health standards.
- **Model Proficiency**: Decision Tree improved notably; Random Forest achieved the highest accuracy.
- **Feature Selection Efficiency**: Minimized redundant predictors to enhance model precision.
- **Implications**: Demonstrated the potential of predictive analytics in influencing health policies and personal health management.

Thank You