STAT 451 Project Report
Group Member: Kate Li, Jing Chang, Qiaoyu(Peter) Bi, Ruicai Cui, Xiaowen(Tommy) Liu

1. **Introduction**

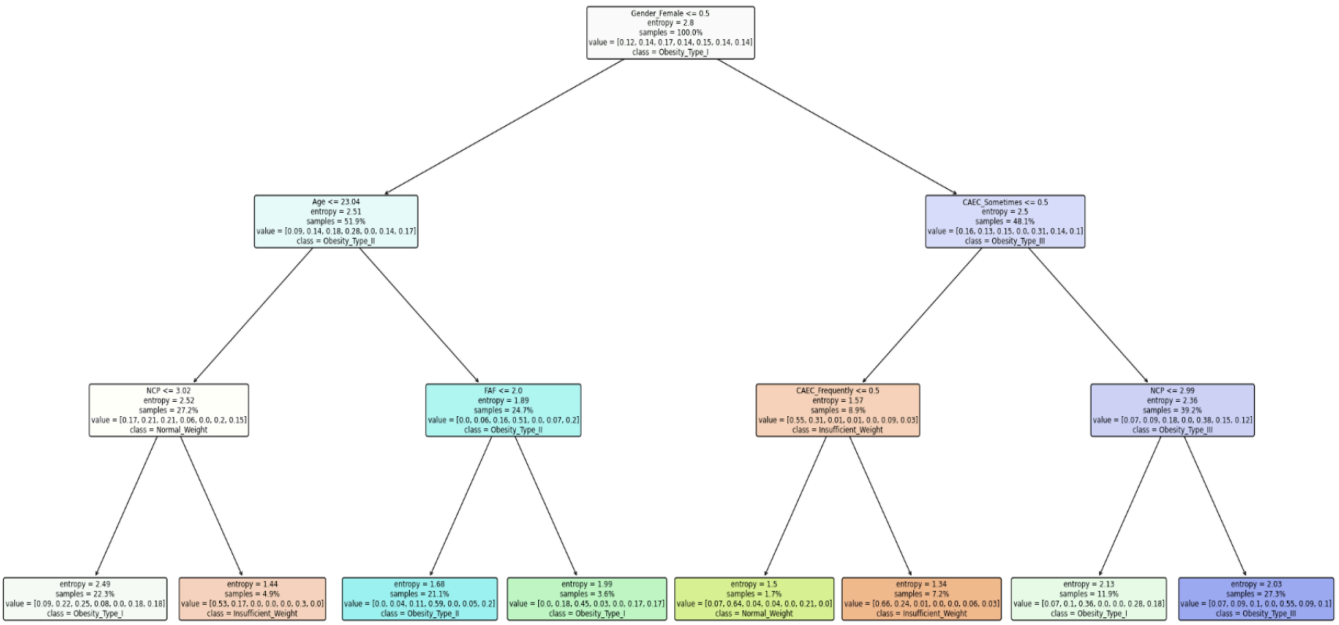
   We developed a machine learning model to classify individuals into obesity levels using physiological and lifestyle data from the UCI Machine Learning Repository's "*Estimation of Obesity Levels Based on Eating Habits and Physical Condition*" dataset, featuring 2,111 records. Numeric variables like Age and CH2O were used as-is, while categorical variables such as Gender and FAVC underwent one-hot encoding with the first category dropped to prevent multicollinearity. CAEC was encoded into ordinal numeric values. The study aimed to address the complex health issue of obesity, influenced by various genetic, environmental, and behavioral factors. We utilized five prediction methods—Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Tree, and Support Vector Machines—to analyze and predict obesity levels. Our analysis focused on optimizing these models to improve prediction accuracy, which was achieved by excluding height and weight to simplify the model while focusing on variables such as age, gender, meal frequency, water intake, physical activity, and snacking habits. The most accurate model, Random Forest, achieved a 73% prediction accuracy, highlighting its effectiveness in distinguishing different obesity categories based on selected features.

2. **Data Analysis**

- Logistic Regression: In our study, we simplified the classification of obesity levels into two categories—'obesity' and 'non-obesity'—to facilitate the use of logistic regression, which is apt for binary outcomes. This reclassification allowed a more straightforward analysis, resulting in a model accuracy of 76.67% on the training data. The strength of Logistic regression is its straightforwardness of implementation, making it accessible to both technical and non-technical audiences. Its weakness is its assumption of linearity between the dependent variable and the independent variables, which can limit its effectiveness in complex scenarios

STAT 451 Project Report
Group Member: Kate Li, Jing Chang, Qiaoyu(Peter) Bi, Ruicai Cui, Xiaowen(Tommy) Liu

- Decision Tree: Set Max Depth to 3 for visualization



The training accuracy of 64%, splits primarily on gender. Females (gender index ≤ 0.5) follow a left sub-tree, splitting further by age, NCP, and FAF. Males proceed to a right sub-tree, splitting on CAEC_sometimes, CAEC_frequently, and NCP. Decision trees offer easy interpretability but are susceptible to overfitting and instability.

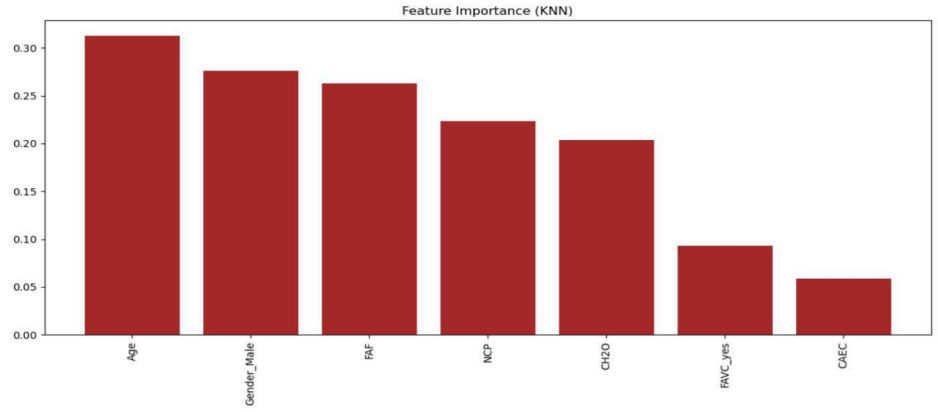
- KNN: Achieved 66% accuracy, excelling in identifying high-risk obesity with a recall of 0.98 but struggled with normal weight categories. It displayed varied precision and F1 scores across categories, indicating better sensitivity in predicting severe obesity types. It is transparent and interpretable, making it easier to understand the influence of each feature. It's computationally efficient and effective for binary classification tasks, but it might not hold for complex relationships.

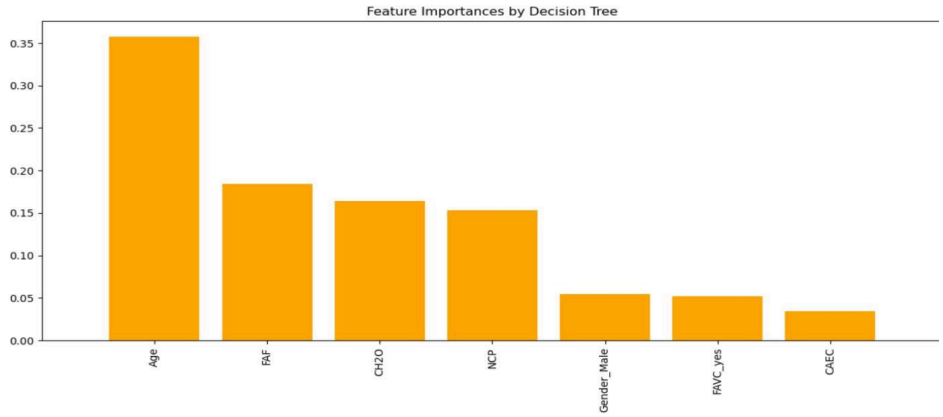
- SVM: Using a Grid Search and cross-validation, optimal SVM parameters were identified, improving the model's accuracy to 68%. The model excels in the 'Obesity_Type_III' category with high recall and F1 scores but shows weaker recall in the 'Overweight_Level_II' category. This improvement post-tuning underscores the effectiveness of the optimization process in enhancing model performance. The strength of the SVM model is its ability to effectively handle high-dimensional data and perform well with a clear margin of separation. Its weakness is its susceptibility to overfitting.

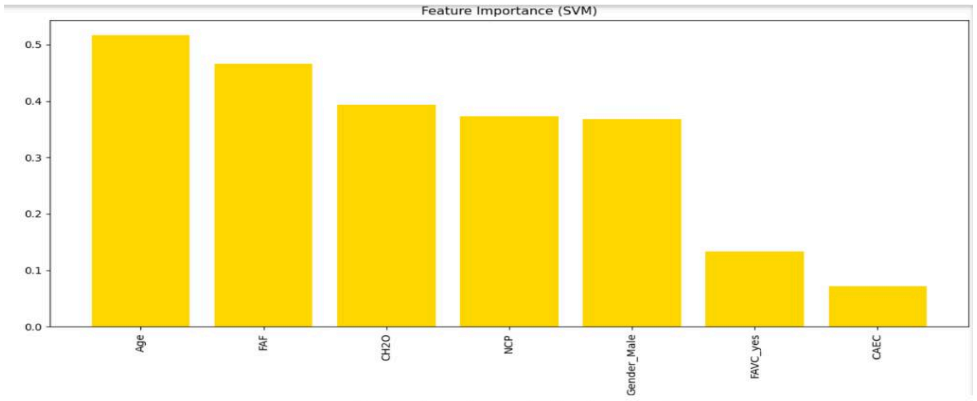
- Random Forest: It outperforms the other models with 73% accuracy, especially in distinguishing the 'Obesity_Type_III' category, with very high precision, recall, and F1 score which are calculated by each category relative to all the rest of categories. Overall, random forests improve prediction accuracy and control overfitting better

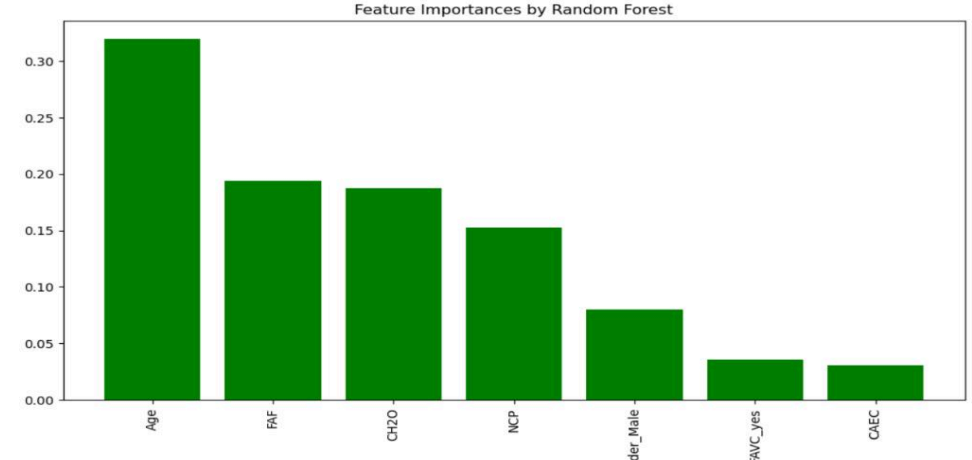
than single trees. However, they are computationally intensive and less interpretable due to ensemble complexity.

3. Feature Importance









The feature importance graphs across different models (KNN, Decision Tree, SVM, and Random Forest) highlight varying impacts of features on obesity level predictions. Age consistently shows high importance across all models, indicating a strong correlation with obesity levels. Other variables like FAF (physical activity frequency) and CH2O (water intake) also show varied importance, emphasizing their roles in obesity prediction but to different extents across models**.**

4. **Conclusion**

Our analysis successfully applied machine learning to predict obesity levels from lifestyle and physiological data. The Random Forest model showed the highest accuracy at 73%. The project's weaknesses include potential overfitting, limited generalizability due to reliance on a single dataset and lack of a deep statistical validation of differences. Future studies could explore the integration of additional behavioral factors and the application of more complex algorithms like deep learning to enhance predictive accuracy and provide more personalized health recommendations.

5. **Contribution**

Member	Proposal	Coding	Presentation	Report
Kate Li	0.7	0.8	1	1
Ruicai Cui	0.8	1	1	1
Qiaoyu(Peter) Bi	1	1	1	1
Jing Chang	0.8	0.8	1	1
Xiaowen(Tommy) Liu	1	0.8	1	1