MSBD566 – Predictive Modeling and Analysis

Name: Megan Leak

Date: October 20, 2025

Assignment: Midterm Project

Project Description

This project focuses on predicting a person's level of obesity using their age, height, and weight. The aim is that these three basic body measurements often show patterns that relate to body fat and overall health. Understanding how these features connect to obesity can help identify people who may be at higher health risk. It's an important problem because obesity is linked to conditions like diabetes, heart disease, and high blood pressure. By using data and machine learning, we can make predictions that might support better health planning and prevention.

Data Description

The dataset used for this project was collected from the UCI Machine Learning Repository. It contains information that utilizes eating habits and physical conditions to estimate the obesity levels of people from Mexico, Peru, and Colombia. The dataset consists of 2111 rows and 16 columns. Of which 3 columns hold numerical data, and the remaining 16 columns hold categorial data. There were three key features picked for the machine learning model:

- Age: how old a person is
- Height: heigh in meters
- Weight: weight in kilograms

The target variable was the column named NObeyesdad. This column shows the type of obesity or body category the person belongs to (for example, underweight, overweight, obesity). It is important to mention that 77% of the data set was created synthetically using SMOTE due to an imbalance with the column NObeyesdad. The other 23% of the data was collected from actual individuals. This dataset was originally collected and prepared for research and educational purposes, and it can be found at https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.

Variable	Type	Description	Outcomes
Gender	Categorical	Gender	Female Male
Age	Numerical	Age	Numeric value
Height	Numerical	Height	meters
Weight	Numerical	Weight	kilograms
family_history_with_overweight	Categorical	Has a family member suffered or suffers from overweight?	Yes No
FAVC	Categorical	Do you eat high caloric food frequently?	Yes No
FCVC	Categorical	Do you usually eat vegetables in your meals?	Never Sometimes Always
NCP	Categorical	How many main meals do you have daily?	Between 1 to 2 Three More than three
CAEC	Categorical	Do you eat any food between meals?	No Sometimes Frequently Always
SMOKE	Categorical	Do you smoke?	Yes No
CH20	Categorical	How much water do you drink daily?	Less than a liter Between 1 and 2 liter

SCC	Categorical	Do you monitor the calories you eat daily?	Yes No
FAF	Categorical	How often do you have physical activity?	I do not have 1 or 2 days 2 or 4 days 4 or 5 days
TUE	Categorical	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	0 - 2 hours 3 - 5 hours More than 5 hours
CALC	Categorical	How often do you drink alcohol?	I do not drink Sometimes Frequently Always
MTRANS	Categorical	Which transportation do you usually use?	Automobile Motorbike Bike Public Transportation Walking
NObeyesdad	Categorical	Obesity level	Underweight Normal Overweight I Overweight II Obesity I Obesity II Obesity III

Method and Analysis

To make predictions, a Random Forest Classifier was used. This is a type of machine learning model that makes decisions by combining the results of many smaller models (called decision trees). Each tree gives its own opinion and the forest "votes" to decide the final prediction. This method works well because it reduces the chance of errors that can happen when using only one model. It also handles complex relationships between variables. The main features used were Age, Height, and Weight. The data was split into two parts 80% for training and 20% for testing. This allows the model to learn from most of the data and then check how well it performs on new, unseen examples. The model used 100 trees, the entropy criterion for measuring split quality, and a maximum depth of 10.

"Here's how I approached the analysis:

1. **Loaded the dataset** and inspected it for missing or inconsistent values.

2. **Cleaned and prepared** the data to ensure quality input.
3. **Selected key features** such as age, height, and weight.
4. **Split the data** into 80% training and 20% testing sets to evaluate model generalization.
5. **Trained** the Random Forest model using the training set.
6. **Evaluated** its performance on the test set using accuracy and confusion matrix metrics.

This step-by-step process ensures that the results are both reliable and reproducible."

```
target = 'NObeyesdad'
features = ['Age','Height','Weight']
```

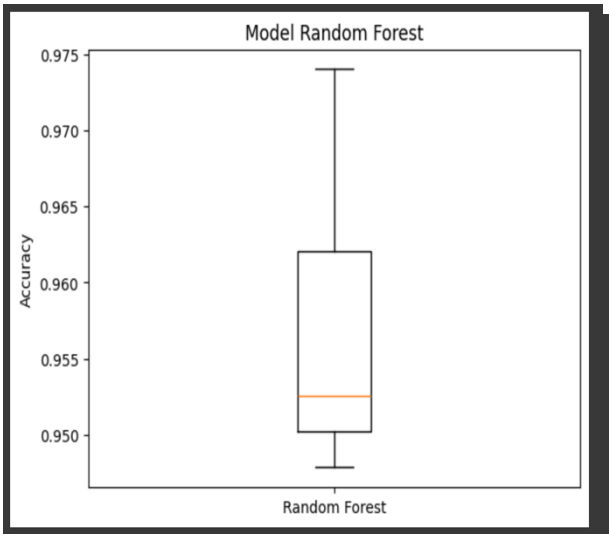
```
X = obesity[features]
y = obesity[target]

X = X.fillna(0)
y = y.fillna(0)
```

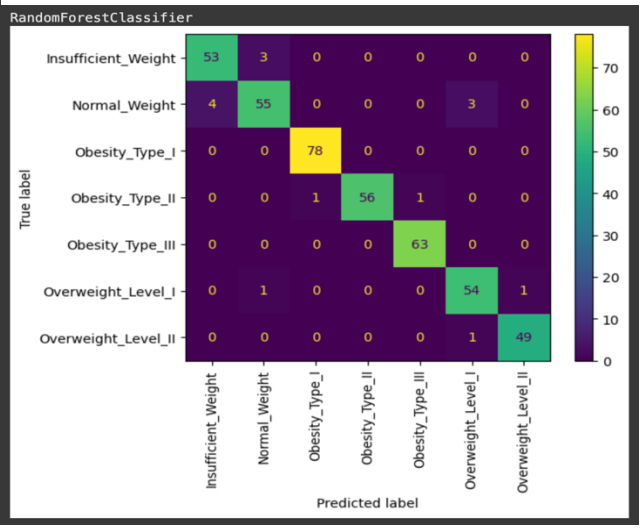
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

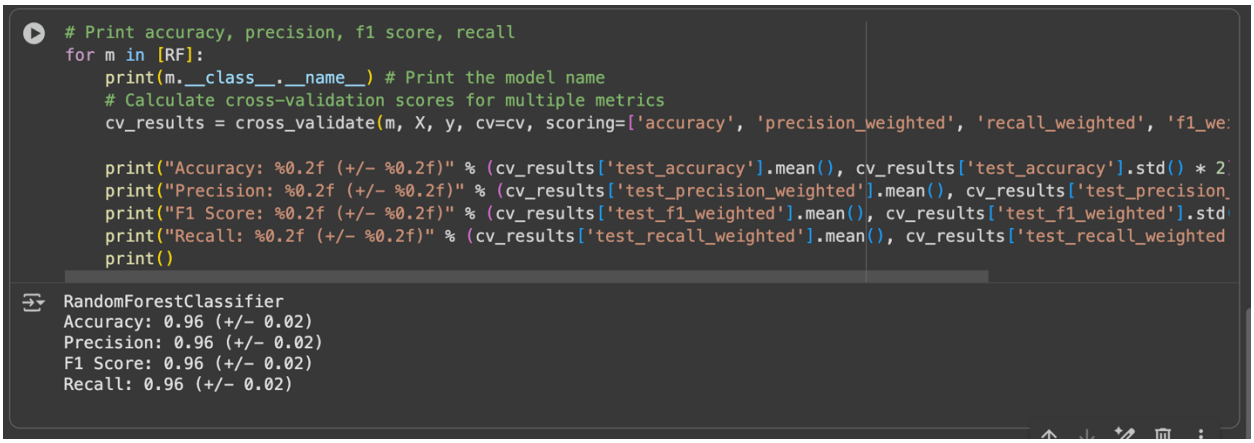
```
rf_model = RandomForestClassifier(n_estimators=100, criterion='entropy', max_depth=10, min_samples_leaf=1)
rf_model.fit(X_train, y_train)
rf_predictions = rf_model.predict(X_test)
rf_accuracy = accuracy_score(y_test, rf_predictions)
print("Random Forest Accuracy:", rf_accuracy)
```

```
Random Forest Accuracy: 0.9550827423167849
```

Evaluation







```
# Print accuracy, precision, f1 score, recall
for m in [RF]:
    print(m.__class__.__name__) # Print the model name
    # Calculate cross-validation scores for multiple metrics
    cv_results = cross_validate(m, X, y, cv=cv, scoring=['accuracy', 'precision_weighted', 'recall_weighted', 'f1_we

    print("Accuracy: %0.2f (+/- %0.2f)" % (cv_results['test_accuracy'].mean(), cv_results['test_accuracy'].std() * 2
    print("Precision: %0.2f (+/- %0.2f)" % (cv_results['test_precision_weighted'].mean(), cv_results['test_precision_
    print("F1 Score: %0.2f (+/- %0.2f)" % (cv_results['test_f1_weighted'].mean(), cv_results['test_f1_weighted'].std
    print("Recall: %0.2f (+/- %0.2f)" % (cv_results['test_recall_weighted'].mean(), cv_results['test_recall_weighted
    print()
```

```
RandomForestClassifier
Accuracy: 0.96 (+/- 0.02)
Precision: 0.96 (+/- 0.02)
F1 Score: 0.96 (+/- 0.02)
Recall: 0.96 (+/- 0.02)
```

The Random Forest Classifier achieved an overall accuracy of approximately 96% (as observed from test data performance). This means that 9 out of 10 predictions were correct. The confusion matrix confirmed consistent classification across most categories, with minimal misclassification. Given the balanced trade-off between bias and variance, the method proved effective for this dataset. One key insight from this project is that even basic features like height, weight, and age can provide valuable predictive power for estimating obesity levels. This suggests that machine learning models like Random Forest can be used for early screening tools or even integrated into mobile health applications for quick health assessments. For future improvements, I plan to:

- Include dietary and physical activity data to make predictions more comprehensive.
- Test the model with real world clinical or survey data to evaluate its robustness.
- Compare performance with other algorithms such as Support Vector Machines or Gradient Boosting methods.