Shady Dajani

Behavioral Data & Individual Health Classification

Problem Statement

Humans have been progressing towards an age where maintaining a healthy lifestyle has become a priority. This can play a crucial role in everyone's life so it is critical we investigate the behavioral factors that can lead to health issues such as high BMI, which is an underlying issue in many health-related problems. Using this data, I will use exploratory analysis and see if there is a relationship between high BMI and the frequency of eating out by conducting A/B tests and utilizing ML models.

Data

The data pulled is from Kaggle and it is structured synthetic and real-world data. The data demonstrates people and their lifestyle behaviors such as fast food consumption. The type of data being used includes both qualitative and quantitative data. For qualitative data, it is nominal information such as having digestive issues, gender and age (as age is not grouped into categorical ranges). The remainder of the data is quantitative with both discrete and continuous data including number of fast food meals per week, daily calories, BMI, physical activity per week, sleep per day, energy (on a scalable score), doctor visits (per year) and an overall health score.

Preprocessing and Processing Steps

Data Cleaning:

Firstly, I checked for missing values in each column and found that there were no missing values, which was not surprising due to the data being synthetic. Although, if missing values were present it would not be an issue because you can drop the rows of missing values using df.dropna(). This would be an important step because it ensures there are no null values in the data leading to issues later on.

Data Filtering & Encoding:

Next, for one of the features specifically gender there were three types of classifications : Male, Female and Other. I kept only Male and Female classifications and encoded to create a binary classification and avoid a dummy variable trap. Encoding helps clean the data to make sure it's ready for models and to avoid errors with training and testing on data with unseen categories.With df=df[df['Gender'].isin(['Male', 'Female'])].copy() rows with Gender_Other were dropped keeping the data consistent. The encoded data is Male = 0 and Female = 1.

Dependent and Independent variables:

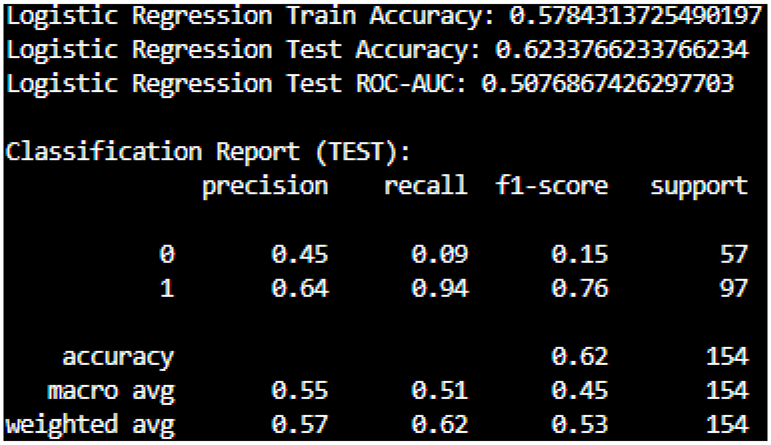
The chosen dependent variable in this analysis is BMI_binary which is derived from BMI. BMI was the chosen dependent variable because typically when people have unhealthy lifestyle habits they tend to have a higher BMI. I created the BMI_binary column from BMI (df['BMI_binary'] = (df['BMI'] > 25).astype(int). This organizes BMI >=25 and lower and higher into groups. According to standard BMI classifications, a BMI of 25 marks the beginning of the overweight range. Using this standard, I can categorize people in the dataset into groups of being “normal” or “overweight”.

The independent variables are the other columns in the dataset that can be considered an influence on a person's BMI but the one I am focusing on is fast food consumption.

Models

The models used in this exploratory analysis are Logistic Regression, Random Forest and Decision Tree. Logistic Regression is a core machine learning algorithm for classification, predicting the probability of a categorical outcome. Random Forest and Decision Tree are tree based models with different uses. Random Forest combines multiple tree-like models to get a better result. Each tree is trained on a random subset of the training data and features. A decision tree is a flowchart-like model to classify or predict outcomes, mimicking human decision-making by splitting data into branches based on feature values.

Logistic Regression:



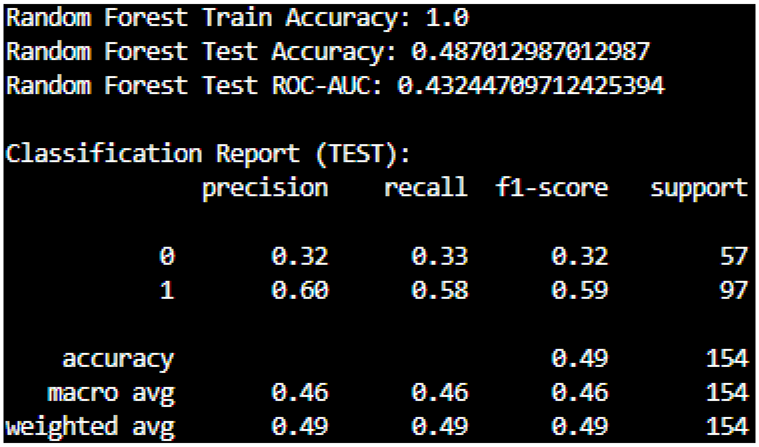
```
Logistic Regression Train Accuracy: 0.5784313725490197
Logistic Regression Test Accuracy: 0.6233766233766234
Logistic Regression Test ROC-AUC: 0.5076867426297703

Classification Report (TEST):
              precision    recall  f1-score   support

           0       0.45      0.09      0.15        57
           1       0.64      0.94      0.76        97

    accuracy                           0.62       154
   macro avg       0.55      0.51      0.45       154
weighted avg       0.57      0.62      0.53       154
```

Random Forest:



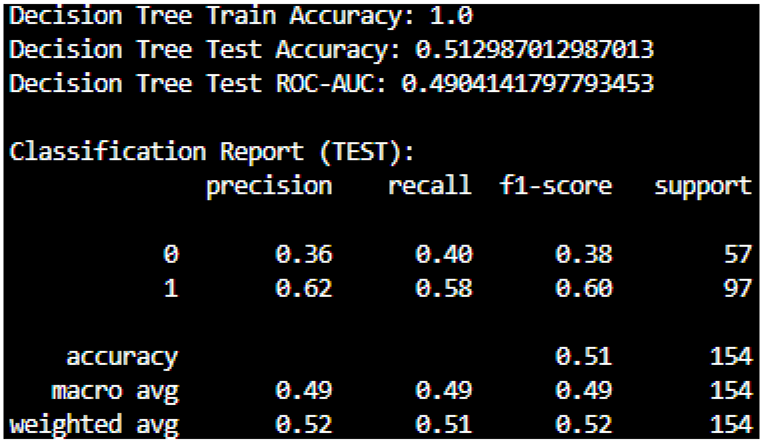
```
Random Forest Train Accuracy: 1.0
Random Forest Test Accuracy: 0.487012987012987
Random Forest Test ROC-AUC: 0.43244709712425394

Classification Report (TEST):
              precision    recall  f1-score   support

           0       0.32      0.33      0.32        57
           1       0.60      0.58      0.59        97

    accuracy                           0.49       154
   macro avg       0.46      0.46      0.46       154
weighted avg       0.49      0.49      0.49       154
```

Decision Tree:



```
Decision Tree Train Accuracy: 1.0
Decision Tree Test Accuracy: 0.512987012987013
Decision Tree Test ROC-AUC: 0.4904141797793453

Classification Report (TEST):
              precision    recall  f1-score   support

           0       0.36      0.40      0.38        57
           1       0.62      0.58      0.60        97

    accuracy                           0.51       154
   macro avg       0.49      0.49      0.49       154
weighted avg       0.52      0.51      0.52       154
```

Training Accuracy:

The Logistic Regression model had a training accuracy of 58%. This lower memorization score suggests that it is a simpler model that avoids memorizing the training data. This also suggests that there is no overfitting and potentially may generalize better to unseen data. Both Random Forest and Decision tree had a testing accuracy of 100% which means the models memorized the data entirely. Since it memorized the data it was likely learning patterns and capturing noise therefore the model is overfitting.

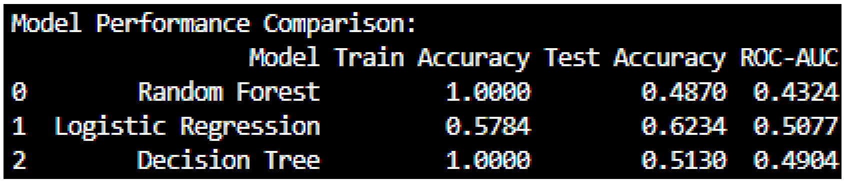
Testing Accuracy:

The Logistic Regression model has the highest testing accuracy at 0.62 , then the Decision Tree model at 0.51, and the lowest being the Random Forest model at 0.49. Although none of the three models performed very strongly, Logistic Regression had the best generalization. The Random Forest and Decision Tree models show a drop of 1.0000 to approximately .50 , going from perfect training accuracy to 50%, which confirms overfitting

AUROC Score:

The AUROC score reflects how well the models distinguish between classes. The Logistic Regression model performs best with a score of 0.51, followed by the Decision Tree at 0.49, and the lowest being the Random Forest at 0.43. Since all scores are mostly near 0.5, it indicates performance similar to random guessing. This implies that none of the models are effective at distinguishing between high BMI and lifestyle factors, but the Logistic Regression model performs slightly better than the other models. AUROC near 0.5 implies that the model has poor ability to distinguish between classes.

Comparison:



```
Model Performance Comparison:
                 Model  Train Accuracy  Test Accuracy  ROC-AUC
0        Random Forest          1.0000         0.4870   0.4324
1  Logistic Regression          0.5784         0.6234   0.5077
2        Decision Tree          1.0000         0.5130   0.4904
```

Results:

Fast food frequency shows high variance in high BMI, indicating it is a weak standalone predictor and must be interpreted alongside lifestyle factors.

Exploratory analysis showed that average BMI does not increase monotonically with fast-food consumption frequency. Significant overlap across groups suggests fast-food intake alone is a weak predictor of BMI and overall health outcomes. This limited separability explains the modest ROC-AUC observed across models and supports the selection of Logistic Regression for its stability and interpretability.

Conclusion

All in all, the results of this project suggest that behavioral factors provide meaningful insights, however they are not sufficient on their own to accurately classify individual health outcomes. In the analysis, I discussed different lifestyle factors that could contribute to a high BMI such as Fast_Food_Meals_Per_Week, Average_Daily_Calories, Physical_Activity_Hours_Per_Week, Sleep_Hours_Per_Day, etc. When I dove into the findings of the 3 machine learning models, I found that the Logistic Regression Model not only has superior generalization performance (modest ROC-AUC), but also had more stability on the data that was given and interpretability compared to the more complex tree-based models. The two tree-based models achieved perfect training accuracy, meaning the data was just memorized. Also because the training data did well and unseen data didn't even surpass .5 test accuracy, this demonstrates significant overfitting and poor ranking performance. There are limitations such as having accurate self-reported behaviors and the absence of biological features which can result in a difference in results. Lastly, this analysis could have been improved by getting a larger dataset, modeling BMI as a continuous variable and exploring more machine learning models capable of capturing non-linear relationships.

Resources

https://www.kaggle.com/datasets/prince7489/fast-food-consumption-and-health-impact-dataset?resource=download