Avengers Report Multi Class Prediction of Obesity Risk

1st Suhai LUO
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
suhailuo@ln.hk

2nd Yuqi GUAN
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
yuqiguan@LN.hk

3rd Xinyu DONG
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
xinyudong@LN.hk

4th Xinghua HUANG
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
xinghuahuang@LN.hk

5th Bowen TANG
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
btang1@LN.hk

6th Linling SHEN
School of Data Science
Lingnan University
Tuen Mun, Hong Kong
linlingshen@LN.hk

***Abstract*—The study utilize estimation of obesity levels data to figure out possible relationships among basic information, physical health indicators and obesity level. Random forest algorithm is emphasized in our multi class prediction question. We draw a conclusion that 'Weight' is the most relative attribute to obesity levels and gain overall 0.8839 prediction accuracy.**

***Index Terms*—random forest, prediction, multi class problem**

I. Background

Obesity often leads to social stigmatization and discrimination, which can negatively impact individuals' mental health and quality of life. Social perceptions of obesity can create psychological distress, including feelings of shame, low self-esteem, and depression [1]. From a health perspective, obesity is a major risk factor for a variety of chronic diseases. The condition increases the likelihood of developing cardiovascular diseases, type 2 diabetes, hypertension, certain cancers, and respiratory issues [2].

The risks associated with obesity are vast and multifaceted. Obesity is one of the leading causes of preventable mortality, with higher risks of heart disease, stroke, and diabetes [3]. Those are the substantial evidences for us to focusing on obesity and its causes.

II. Data Collection & Description

The current challenge focuses on predicting obesity risk, a complex problem that involves understanding how various features correlate with health outcomes. This setup allows data miners to practice identifying the most relevant features, managing imbalanced data, and deploying techniques to avoid over fitting all critical skills in data mining.

This research was conducted under the Data Mining Course delivered by Professor DONG, Division of Artificial Intelligence, School of Data Science, Lingnan University.

The data consist of the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition , data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records [4]. The synthetic nature of the dataset also ensures that participants can freely experiment without concerns over data privacy or ethical constraints, making it an ideal sandbox for testing new ideas and algorithms.

Data is available in CSV format. The attributes obtained : Gender, Age, Height, Weight, family history with overweight, MTRANS, and Physical Health Indicators. The data contains both numerical data and continuous data, which can be used for analysis based on algorithms of classification, prediction, segmentation and association.

The variable to be predict is 'NObesity' and its values are shown in Table I below:

TABLE I
NObesity Classification

Value	Measurement
Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight	25.0 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

The Physical Health Indicators are specified: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC).

```
test.head
       id  Gender        Age  ...       TUE       CALC                MTRANS
0  20758    Male  26.899886  ...  0.000000  Sometimes  Public_Transportation
1  20759  Female  21.000000  ...  0.000000  Sometimes  Public_Transportation
2  20760  Female  26.000000  ...  0.250502  Sometimes  Public_Transportation
3  20761    Male  20.979254  ...  0.000000  Sometimes  Public_Transportation
4  20762  Female  26.000000  ...  0.741069  Sometimes  Public_Transportation
```

Fig. 1. Data exploration using head().

```
train.CALC
 CALC
Sometimes     15066
no             5163
Frequently      529
Name: count, dtype: int64

test.CALC
 CALC
Sometimes     9979
no            3513
Frequently     346
Always           2
Name: count, dtype: int64
```

Fig. 2. Different values between train and test sets.

The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS) [5].

III. Data Cleaning & Preprocessing

Steps by steps, we do the followings to preprocessing data in order to meet the prerequisite of model training.

- Load the Data

We import the necessary libraries pandas, sklearn. Then load both train.csv, test.csv files into pandas Data Frames.

- Initial Data Exploration

Use info(), describe(), head() to check for column data types and null values and understand the numerical data distribution as Fig. 1 showed.

- Handle Outliers & Missing Values

If there are many outliers or missing values, consider fulfilling or dropping the column. Remove attributes if they are irrelative to prediction.

- Data Type Conversion

Ensure that categorical features are of type category or object then convert them into numeric for further modeling.

- Align Train and Test Sets

Training sets and testing sets might have different values as Fig. 2 showed so we need to ensure that both train and test datasets have the same features eventually, especially after encoding categorical features.

```
val_accuracy
 Validation Accuracy: 0.8839113680154143
```

Fig. 3. Random forest prediction accuracy.

- Feature Engineering

Perhaps the attributes are related so that we can combine several of them to derive a new variable. Typically we can assemble the year, month, and day into a date time column.

IV. Methods Applying

Random forest is a suitable choice for these two datasets due to several key advantages that it offers, especially for tabular data with both numeric and categorical features. The inherit features of random forest include robustness to over fitting and feature importance. Random forest, as an ensemble learning method, combines multiple decision trees built on different subsets of the data. This reduces the risk of over fitting, making it more generalizable to unseen data compared to a single decision tree. It also provides feature importance scores, which can help in understanding which features contribute the most to predictions. This is useful for interpreting model results and refining feature engineering [6].

V. Model Training & Result Interpretation

To apply random forest in our study, we can do the following process assuming that data preprocessing is already completed.

- Load the Preprocessed Data
- Separate Features and Target Variable

Define the target variable (y) and feature set (X) from the training data. Ensure the test data has the same features.

- Split the Training Set

Additional validation is needed within the training set.

- Initialize the Random Forest Model

Choose RandomForestClassifier for classification tasks or RandomForestRegressor for regression tasks.

- Make Predictions

Fit the model on the training set and then predict on the validation or test set.

- Evaluate the Model

Calculate accuracy score to evaluate random forest effect. The result predicts the possible obesity type of each person according to their basic conditions. 'Weight' is the most relative attribute to obesity levels as Fig. 4 showed. 'Obesity Type III' occupies the most while others obtains separately.

VI. Model Comparison

Random Forests, Decision Trees, and Neural Networks are popular machine learning models used for various predictive tasks. Each has its unique strengths and weaknesses, making them suitable for different types of problems and datasets.

```
feature_importances
 Age                0.094866
Height              0.090485
Weight              0.340096
FCVC                0.075388
NCP                 0.033817
CH2O                0.047031
FAF                 0.039251
TUE                 0.046253
Gender_Female       0.043587
Gender_Male         0.035888
dtype: float64
```

Fig. 4. Obesity top 10 feature importances.

Decision Trees are simple, intuitive models that split data into branches to make predictions based on feature values. They create a tree-like structure where each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a class label or regression value.

Random Forests are an ensemble learning method that builds multiple decision trees using random subsets of data and features. The final prediction is made by aggregating the predictions of all individual trees.

Neural Networks are computational models inspired by the human brain's neural networks. They consist of layers of interconnected nodes (neurons) that can learn complex patterns through training on large amounts of data [7].

We chose random forest on this task because of its balance of accuracy and over fitting control with lower computational cost.

VII. Conclusion

As Fig. 5 mentioned for this training task, we focused on building a predictive model using the random forest algorithm on two datasets going through all steps led by data mining process. The prediction accuracy reached 0.8839.

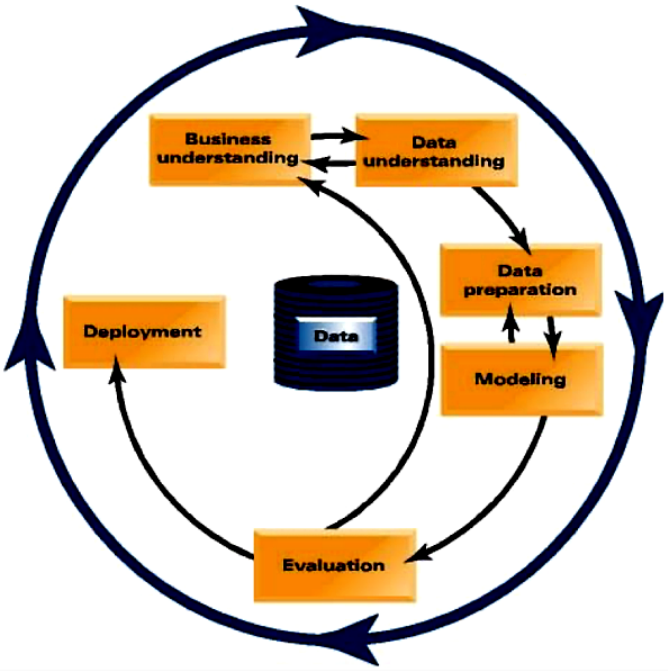

Fig. 5. Data Mining Process.

References

[1] Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributable to obesity: Payer-and service-specific estimates. Health Affairs, 28(5), w822-w831.

[2] Hruby, A., & Hu, F. B. (2015). The epidemiology of obesity: A big picture. PharmacoEconomics, 33(7), 673-689.

[3] Luppino, F. S., de Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W., & Zitman, F. G. (2010). Overweight, obesity, and depression: A systematic review and meta-analysis of longitudinal studies. Archives of General Psychiatry, 67(3), 220-229.

[4] Obesity or CVD risk. [Online]. Available: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster

[5] Multi-Class Prediction of Obesity Risk. [Online]. Available: https://www.kaggle.com/competitions/playground-series-s4e2/data

[6] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. BMC Bioinformatics, 8(1), 1-21.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Appendix

Fig. 6. Python Code: main.py