Peter the Great
Saint-Petersburg Polytechnic University

Multi-Class Prediction of Obesity Risk

Результаты применения модели Random Forest для предсказания риска ожирения у людей на основе различных факторов

Презентацию подготовила Бабахина Софья

POLYTECH

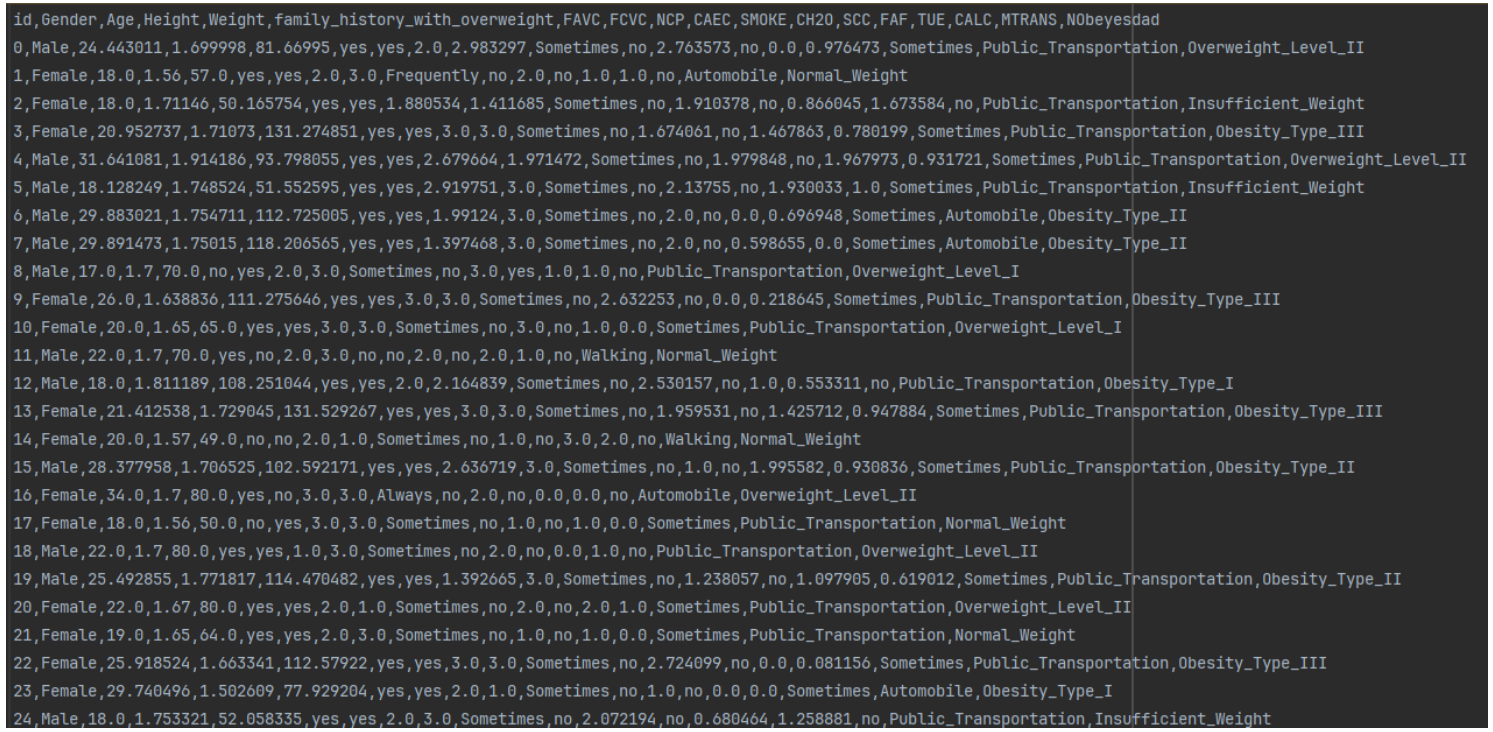
Peter the Great
Saint-Petersburg Polytechnic University

Постановка задачи

Цель: предсказать риск ожирения у людей на основе различных факторов.

Модель: Random Forest

Должно быть выполнено:
1) Исследовательский анализ данных (EDA)
2) Предварительная обработка данных
3) Построение и оценка модели.



```
id,Gender,Age,Height,Weight,family_history_with_overweight,FAVC,FCVC,NCP,CAEC,SMOKE,CH2O,SCC,FAF,TUE,CALC,MTRANS,NObeyesdad
0,Male,24.443011,1.699998,81.66995,yes,yes,2.0,2.983297,Sometimes,no,2.763573,no,0.0,0.976473,Sometimes,Public_Transportation,Overweight_Level_II
1,Female,18.0,1.56,57.0,yes,yes,2.0,3.0,Frequently,no,2.0,no,1.0,1.0,no,Automobile,Normal_Weight
2,Female,18.0,1.71146,50.165754,yes,yes,1.880534,1.411685,Sometimes,no,1.910378,no,0.866045,1.673584,no,Public_Transportation,Insufficient_Weight
3,Female,20.952737,1.71073,131.274851,yes,yes,3.0,3.0,Sometimes,no,1.674061,no,1.467863,0.780199,Sometimes,Public_Transportation,Obesity_Type_III
4,Male,31.641081,1.914186,93.798055,yes,yes,2.679664,1.971472,Sometimes,no,1.979848,no,1.967973,0.931721,Sometimes,Public_Transportation,Overweight_Level_II
5,Male,18.128249,1.748524,51.552595,yes,yes,2.919751,3.0,Sometimes,no,2.13755,no,1.930033,1.0,Sometimes,Public_Transportation,Insufficient_Weight
6,Male,29.883021,1.754711,112.725005,yes,yes,1.99124,3.0,Sometimes,no,2.0,no,0.0,0.696948,Sometimes,Automobile,Obesity_Type_II
7,Male,29.891473,1.75015,118.206565,yes,yes,1.397468,3.0,Sometimes,no,2.0,no,0.598655,0.0,Sometimes,Automobile,Obesity_Type_II
8,Male,17.0,1.7,70.0,no,yes,2.0,3.0,Sometimes,no,3.0,yes,1.0,1.0,no,Public_Transportation,Overweight_Level_I
9,Female,26.0,1.638836,111.275646,yes,yes,3.0,3.0,Sometimes,no,2.632253,no,0.0,0.218645,Sometimes,Public_Transportation,Obesity_Type_III
10,Female,20.0,1.65,65.0,yes,yes,3.0,3.0,Sometimes,no,3.0,no,1.0,0.0,Sometimes,Public_Transportation,Overweight_Level_I
11,Male,22.0,1.7,70.0,yes,no,2.0,3.0,no,no,2.0,no,2.0,1.0,no,Walking,Normal_Weight
12,Male,18.0,1.811189,108.251044,yes,yes,2.0,2.164839,Sometimes,no,2.530157,no,1.0,0.553311,no,Public_Transportation,Obesity_Type_I
13,Female,21.412538,1.729045,131.529267,yes,yes,3.0,3.0,Sometimes,no,1.959531,no,1.425712,0.947884,Sometimes,Public_Transportation,Obesity_Type_III
14,Female,20.0,1.57,49.0,no,no,2.0,1.0,Sometimes,no,1.0,no,3.0,2.0,no,Walking,Normal_Weight
15,Male,28.377958,1.706525,102.592171,yes,yes,2.636719,3.0,Sometimes,no,1.0,no,1.995582,0.930836,Sometimes,Public_Transportation,Obesity_Type_II
16,Female,34.0,1.7,80.0,yes,no,3.0,3.0,Always,no,2.0,no,0.0,0.0,no,Automobile,Overweight_Level_II
17,Female,18.0,1.56,50.0,no,yes,3.0,3.0,Sometimes,no,1.0,no,1.0,0.0,Sometimes,Public_Transportation,Normal_Weight
18,Male,22.0,1.7,80.0,yes,yes,1.0,3.0,Sometimes,no,2.0,no,0.0,1.0,no,Public_Transportation,Overweight_Level_II
19,Male,25.492855,1.771817,114.470482,yes,yes,1.392665,3.0,Sometimes,no,1.238057,no,1.097905,0.619012,Sometimes,Public_Transportation,Obesity_Type_II
20,Female,22.0,1.67,80.0,yes,yes,2.0,1.0,Sometimes,no,2.0,no,2.0,1.0,Sometimes,Public_Transportation,Overweight_Level_II
21,Female,19.0,1.65,64.0,yes,yes,2.0,3.0,Sometimes,no,1.0,no,1.0,0.0,Sometimes,Public_Transportation,Normal_Weight
22,Female,25.918524,1.663341,112.57922,yes,yes,3.0,3.0,Sometimes,no,2.724099,no,0.0,0.081156,Sometimes,Public_Transportation,Obesity_Type_III
23,Female,29.740496,1.502609,77.929204,yes,yes,2.0,1.0,Sometimes,no,1.0,no,0.0,0.0,Sometimes,Automobile,Obesity_Type_I
24,Male,18.0,1.753321,52.058335,yes,yes,2.0,3.0,Sometimes,no,2.072194,no,0.680464,1.258881,no,Public_Transportation,Insufficient_Weight
```

data.csv

POLYTECH

Peter the Great
Saint-Petersburg Polytechnic University

Принцип работы метода Random Forest

Принцип работы алгоритма:

Random forest («случайный лес») — это алгоритм машинного обучения, который состоит из множества отдельных решающих деревьев, то есть из независимых моделей.

Шаги:

1. Подготовка обучающей выборки для одного дерева
2. Обучение дерева
3. Ансамбль
4. Предсказание

Оценка результатов алгоритма:

Confusion_matrix и **Classification_report** — это инструменты, используемые для оценки качества классификационных моделей в машинном обучении.

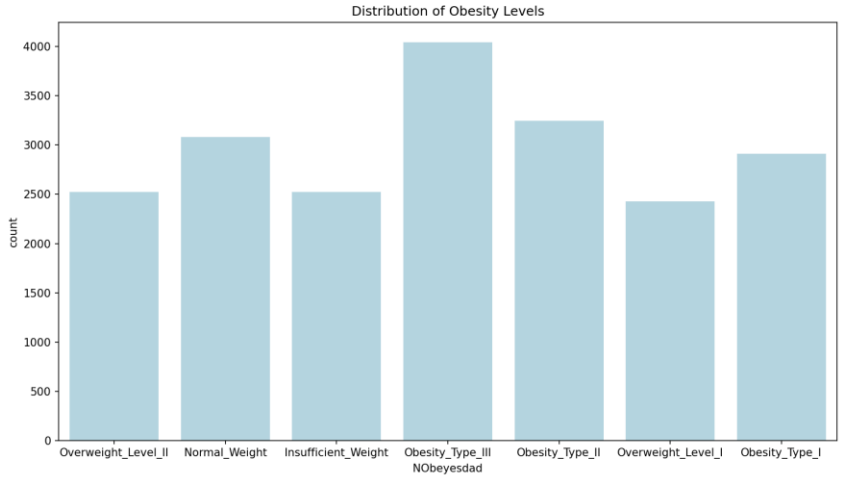
- **Confusion Matrix** — это таблица, которая показывает количество правильных и неправильных предсказаний модели по каждому классу. Она помогает визуализировать производительность модели.
- **Classification Report** — это более подробный отчет о производительности модели, который включает в себя несколько метрик.

POLYTECH

Peter the Great
Saint-Petersburg Polytechnic University

Процесс и результаты

Исследовательский анализ данных



Предварительная обработка данных

```
data['Gender'] = data['Gender'].map({'Male': 0, 'Female': 1})
data['family_history_with_overweight'] = data['family_history_with_overweight'].map({'yes': 1, 'no': 0})
data['FAVC'] = data['FAVC'].map({'yes': 1, 'no': 0})
data['NObeyesdad'] = data['NObeyesdad'].map({
    'Insufficient_Weight': 0,
    'Normal_Weight': 1,
    'Overweight_Level_I': 2,
    'Overweight_Level_II': 3,
    'Obesity_Type_I': 4,
    'Obesity_Type_II': 5,
    'Obesity_Type_III': 6
})
data['CALC'] = data['CALC'].map({'Sometimes': 1, 'no': 0, 'Frequently': 2})
data['CAEC'] = data['CALC'].map({'Sometimes': 1, 'no': 0, 'Frequently': 2, 'Always': 3})
data['SMOKE'] = data['SMOKE'].map({'yes': 1, 'no': 0})
data['SCC'] = data['SMOKE'].map({'yes': 1, 'no': 0})
data['MTRANS'] = data['MTRANS'].map({'Public_Transportation': 0, 'Automobile': 1, 'Walking': 2, 'Motorbike': 3, 'Bike': 4})
```

- **Исследовательский анализ данных (EDA)**: Загружаем данные, проверяем на пропуски и визуализируем распределение целевой переменной.
- **Предварительная обработка данных**: Преобразуем категориальные переменные в числовые и разделяем данные на признаки и целевую переменную.
- **Построение и оценка модели**: Обучаем модель Random Forest и оцениваем её с помощью confusion_matrix и classification_report.

POLYTECH

Peter the Great
Saint-Petersburg Polytechnic University

Процесс и результаты

```
confusion_matrix
 [[ 688   63    3    1    0    0    0]
 [  38  794   64   11    0    0    0]
 [   4   66  561   84   18    0    0]
 [   0   14   76  603   64    6    0]
 [   2    1   27   58  742   25    3]
 [   0    0    0    4   16  983    2]
 [   1    0    1    0    0    1 1204]]
```

```
classification_report
              precision    recall  f1-score   support

           0       0.94      0.91      0.92       755
           1       0.85      0.88      0.86       907
           2       0.77      0.77      0.77       733
           3       0.79      0.79      0.79       763
           4       0.88      0.86      0.87       858
           5       0.97      0.98      0.97      1005
           6       1.00      1.00      1.00      1207

    accuracy                           0.90      6228
   macro avg       0.88      0.88      0.88      6228
weighted avg       0.90      0.90      0.90      6228
```

1. По результатам в матрице ошибок видим (исходя из блоков TN и TP матрицы), что модель довольно точно описывает данные.
2. Классификационный отчёт: видим, что получены высокие показатели precision, recall и f1 для каждого из классов 0-6 и accuracy, что говорит о высокой точности результатов применения модели Random Forest для предсказания риска ожирения у людей на основе различных факторов.

POLYTECH