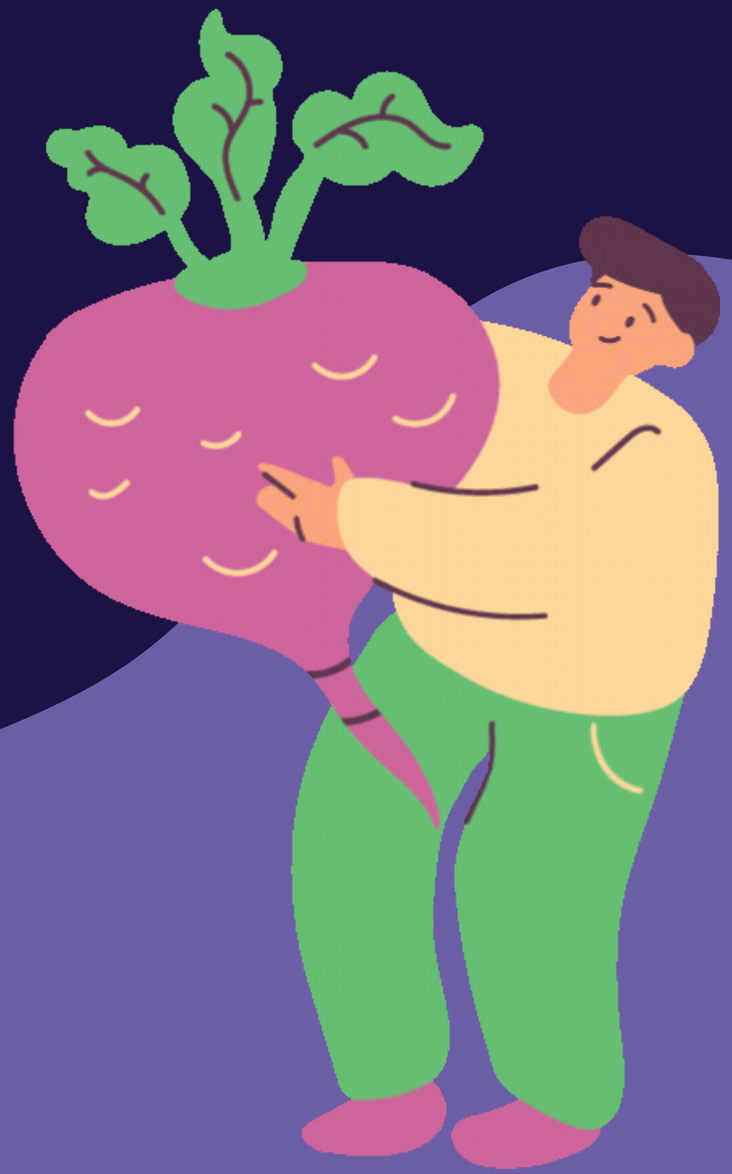
Universidade Estadual do Oeste do Paraná
Disciplina de Inteligência Artificial
Docentes: Huei D. Lee e Newton Spolaor

Checkpoint 2

Mineração de dados

Projeto 4



Discentes:
Isabela Loebel
Nickolas Crema

Base escolhida

Estimation of Obesity Levels Based On Eating Habits and Physical Condition;

(Estimativa dos níveis de obesidade com base nos hábitos alimentares e na condição física - Tradução livre)

- *Fabio Mendoza Palechor e Alexis De la Hoz Manotas, 2019;*
- *Países do estudo: México, Peru e Colômbia;*
- *16 atributos, 1 classe:*
  - *8 atributos numéricos;*
  - *8 atributos categóricos.*
- *2.116 casos;*

Disponível em: <https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

Atributos

- Gênero : categórico;
- Idade : numérico;
- Peso : numérico;
- Altura : numérico;
- Histórico de sobrepeso familiar : categórico;
- Come comida muito calórica: categórico;
- Come vegetais nas refeições : numérico;
- Quantidade de refeições diárias: numérico;
- Come entre as refeições: categórico;
- Fuma : categórico;
- Consumo de água diário : numérico;
- Monitora calórias : categórico;
- Atividades físicas semanais: numérico;
- Tempo gasto com dispositivos tecnológicos : numérico;
- Frequência de consumo de alcool: categórico;
- Meio de transporte: categórico.

Proposta

- Estimar o nível de obesidade de um indivíduo com base na sua condição física e hábitos saudáveis;
- Algoritmos de aprendizado escolhidos:
  - J48 Tree;
  - Random Forest.

Primeiro contato com os dados

- Classe bem distribuida;
- Dataset balanceado por meio de SMOTE pelo criador.

Selected attribute

Name: obesity_level Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	Normal_Weight	287	287
2	Overweight_Level_I	290	290
3	Overweight_Level_II	290	290
4	Obesity_Type_I	351	351
5	Insufficient_Weight	272	272
6	Obesity_Type_II	297	297
7	Obesity_Type_III	324	324

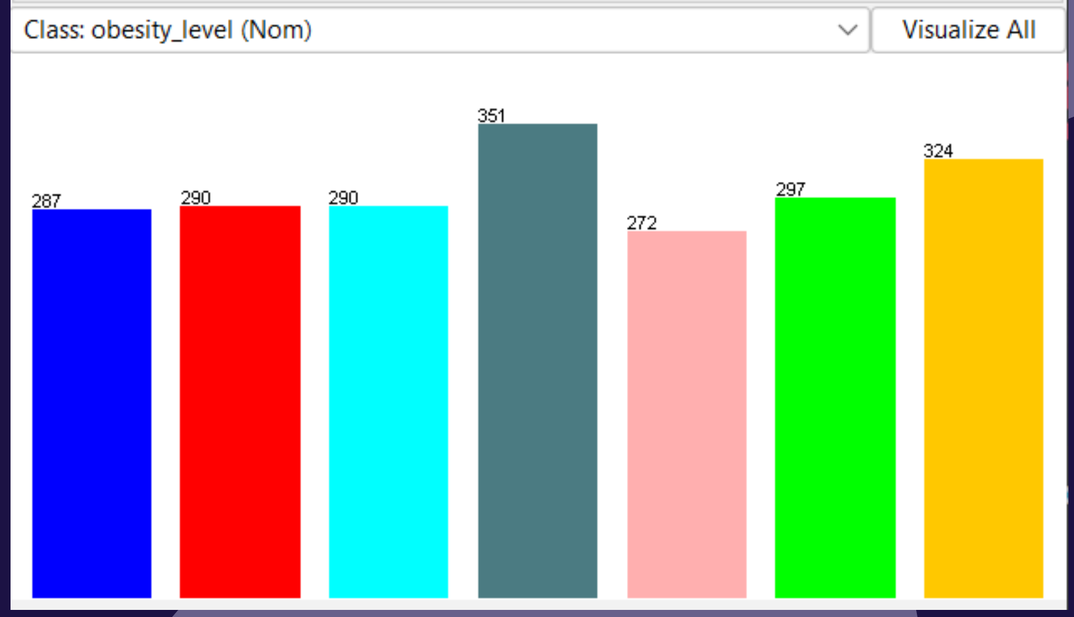


Figura 1 - Distribuição da classe. Fonte: Autores.

Treino e teste

Modelo J48 Tree em cima dos dados sem pré-processamento.

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances         576               90.9953 %
Incorrectly Classified Instances        57                9.0047 %
Kappa statistic                          0.8947
Mean absolute error                      0.029
Root mean squared error                  0.1541
Relative absolute error                 11.8375 %
Root relative squared error             44.0587 %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,815    0,039    0,781      0,815    0,798      0,763    0,930     0,741     Normal_Weight
                 0,854    0,024    0,843      0,854    0,848      0,826    0,937     0,805     Overweight_Level_I
                 0,904    0,015    0,904      0,904    0,904      0,889    0,961     0,888     Overweight_Level_II
                 0,946    0,008    0,963      0,946    0,955      0,945    0,976     0,938     Obesity_Type_I
                 0,870    0,013    0,905      0,870    0,887      0,872    0,990     0,881     Insufficient_Weight
                 0,967    0,006    0,967      0,967    0,967      0,961    0,990     0,950     Obesity_Type_II
                 0,990    0,002    0,990      0,990    0,990      0,988    0,994     0,981     Obesity_Type_III
Weighted Avg.    0,910    0,015    0,911      0,910    0,910      0,896    0,969     0,887

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  75   8   2   0   7   0   0 |   a = Normal_Weight
  10  70   2   0   0   0   0 |   b = Overweight_Level_I
   1   5  75   2   0   0   0 |   c = Overweight_Level_II
   0   0   4 105   0   2   0 |   d = Obesity_Type_I
  10   0   0   0  67   0   0 |   e = Insufficient_Weight
   0   0   0   2   0  88   1 |   f = Obesity_Type_II
   0   0   0   0   0   1  96 |   g = Obesity_Type_III
```

Figura 2 - Resultado do modelo J48 Tree em cima dos dados sem pré-processamento. Fonte: Autores.

Treino e teste

Modelo Random Forest em cima dos dados sem pré-processamento

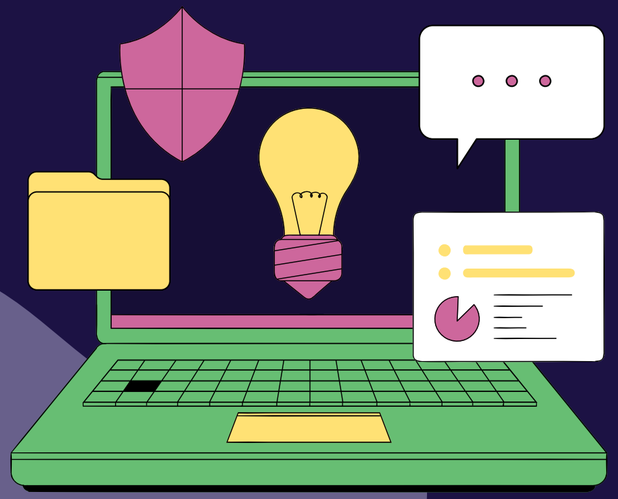
Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances         602               95.1027 %
Incorrectly Classified Instances        31                4.8973 %
Kappa statistic                          0.9427
Mean absolute error                      0.0492
Root mean squared error                  0.1228
Relative absolute error                 20.1318 %
Root relative squared error             35.1136 %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,935    0,026    0,860      0,935    0,896      0,878    0,992     0,954     Normal_Weight
                 0,854    0,013    0,909      0,854    0,881      0,864    0,994     0,973     Overweight_Level_I
                 0,928    0,005    0,963      0,928    0,945      0,937    0,997     0,981     Overweight_Level_II
                 0,982    0,004    0,982      0,982    0,982      0,978    1,000     0,999     Obesity_Type_I
                 0,961    0,004    0,974      0,961    0,967      0,963    0,999     0,992     Insufficient_Weight
                 0,989    0,004    0,978      0,989    0,984      0,981    1,000     0,999     Obesity_Type_II
                 0,990    0,002    0,990      0,990    0,990      0,988    1,000     1,000     Obesity_Type_III
Weighted Avg.    0,951    0,008    0,952      0,951    0,951      0,943    0,998     0,986

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  86   4   0   0   2   0   0 |   a = Normal_Weight
  10  70   2   0   0   0   0 |   b = Overweight_Level_I
   1   3  77   2   0   0   0 |   c = Overweight_Level_II
   0   0   1 109   0   1   0 |   d = Obesity_Type_I
   3   0   0   0  74   0   0 |   e = Insufficient_Weight
   0   0   0   0   0  90   1 |   f = Obesity_Type_II
   0   0   0   0   0   1  96 |   g = Obesity_Type_III
```

Figura 3 - Resultado do modelo Random Forest em cima dos dados sem pré-processamento. Fonte: Autores.

Seleção de atributos

Seleção de Atributos com CfsSubsetEval.



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 81
        Merit of best subset found:    0.496

Attribute Subset Evaluator (supervised, Class (nominal): 17 obesity_level):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,2,4,5,6,7,8,9,11,14 : 10
                     Gender
                     Age
                     Weight
                     family_history_with_overweight
                     eat_high_caloric_food_frequently
                     eat_vegetables_in_meals
                     many_main_meals_daily
                     eat_food_between_meals
                     water_drink_daily
                     time_spending_with_technological_devices
```

Figura 4 - Seleção de atributos com CfsSubsetEval. Fonte: Autores.

Treino e teste

Modelo J48 Tree com CfsSubsetEval.

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         536               84.6761 %
Incorrectly Classified Instances        97               15.3239 %
Kappa statistic                          0.8209
Mean absolute error                      0.0496
Root mean squared error                  0.1968
Relative absolute error                 20.2652 %
Root relative squared error             56.2938 %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,641    0,035    0,756      0,641    0,694      0,650    0,869     0,636     Normal_Weight
                 0,805    0,033    0,786      0,805    0,795      0,764    0,938     0,721     Overweight_Level_I
                 0,687    0,035    0,750      0,687    0,717      0,677    0,924     0,685     Overweight_Level_II
                 0,883    0,033    0,852      0,883    0,867      0,839    0,955     0,853     Obesity_Type_I
                 0,935    0,034    0,791      0,935    0,857      0,839    0,958     0,714     Insufficient_Weight
                 0,967    0,007    0,957      0,967    0,962      0,955    0,990     0,949     Obesity_Type_II
                 0,990    0,002    0,990      0,990    0,990      0,988    0,994     0,981     Obesity_Type_III
Weighted Avg.    0,847    0,025    0,845      0,847    0,844      0,820    0,948     0,799

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 59  9  5  0 19  0  0 |  a = Normal_Weight
 10 66  5  1  0  0  0 |  b = Overweight_Level_I
  4  7 57 14  0  1  0 |  c = Overweight_Level_II
  0  2  9 98  0  2  0 |  d = Obesity_Type_I
  5  0  0  0 72  0  0 |  e = Insufficient_Weight
  0  0  0  2  0 88  1 |  f = Obesity_Type_II
  0  0  0  0  0  1 96 |  g = Obesity_Type_III
```

Figura 5 - Resultado do modelo J48 Tree com CfsSubsetEval. Fonte: Autores.

Treino e teste

Modelo Random Forest com CfsSubsetEval.

Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances         579               91.4692 %
Incorrectly Classified Instances        54                8.5308 %
Kappa statistic                          0.9002
Mean absolute error                      0.0539
Root mean squared error                  0.1422
Relative absolute error                 22.0301 %
Root relative squared error             40.6742 %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,891    0,043    0,781      0,891    0,832      0,804    0,980     0,864     Normal_Weight
                 0,793    0,015    0,890      0,793    0,839      0,818    0,977     0,930     Overweight_Level_I
                 0,807    0,018    0,870      0,807    0,838      0,815    0,988     0,933     Overweight_Level_II
                 0,973    0,011    0,947      0,973    0,960      0,951    0,999     0,994     Obesity_Type_I
                 0,922    0,005    0,959      0,922    0,940      0,933    0,998     0,986     Insufficient_Weight
                 0,989    0,006    0,968      0,989    0,978      0,975    1,000     0,999     Obesity_Type_II
                 0,990    0,002    0,990      0,990    0,990      0,988    1,000     1,000     Obesity_Type_III
Weighted Avg.    0,915    0,014    0,917      0,915    0,914      0,901    0,992     0,959

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  82   3   4   0   3   0   0 |   a = Normal_Weight
  13  65   4   0   0   0   0 |   b = Overweight_Level_I
   4   5  67   6   0   1   0 |   c = Overweight_Level_II
   0   0   2 108   0   1   0 |   d = Obesity_Type_I
   6   0   0   0  71   0   0 |   e = Insufficient_Weight
   0   0   0   0   0  90   1 |   f = Obesity_Type_II
   0   0   0   0   0   1  96 |   g = Obesity_Type_III
```

Figura 6 - Resultado do modelo Random Forest com CfsSubsetEval. Fonte: Autores.

Cálculo IMC

$$IMC = \frac{PESO}{ALTURA * ALTURA}$$

Selected attribute

Name: IMC | Type: Numeric

Missing: 0 (0%) | Distinct: 1968 | Unique: 1890 (90%)

Statistic	Value
Minimum	12.999
Maximum	50.812
Mean	29.7
StdDev	8.011

Class: obesity_level (Nom) | Visualize All

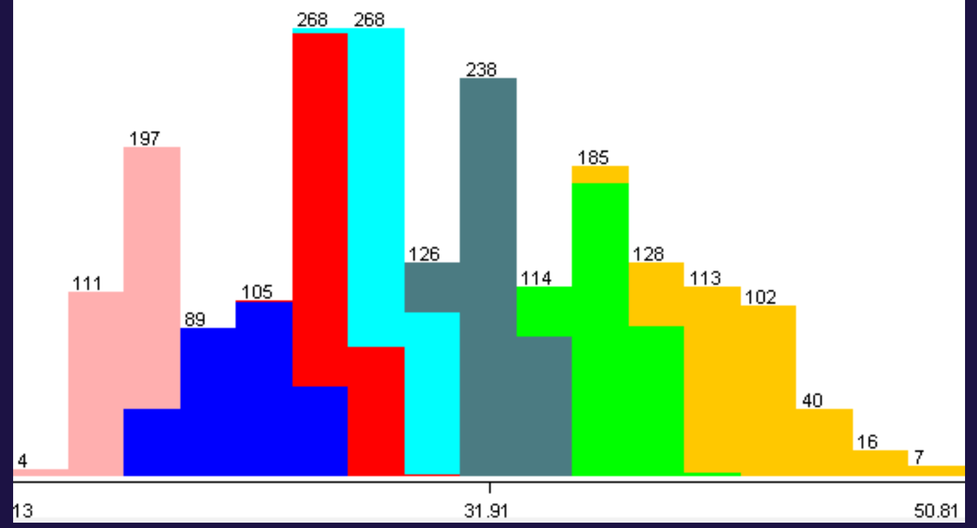
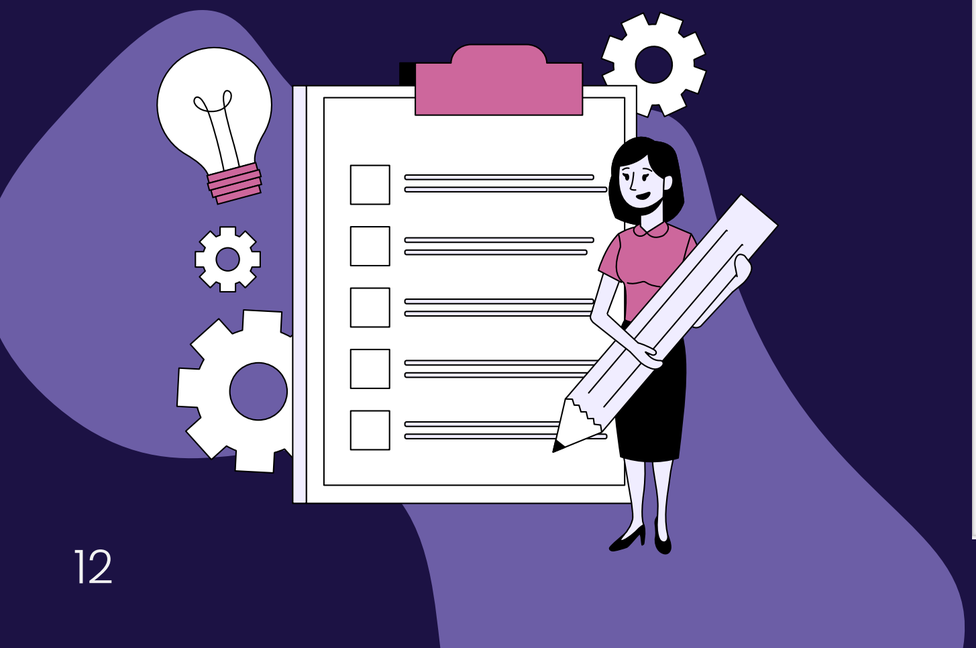


Figura 7 - Novo atributo criado com base no IMC. Fonte: Autores.

Treino e teste

Modelo J48 Tree com CfsSubsetEval com IMC.



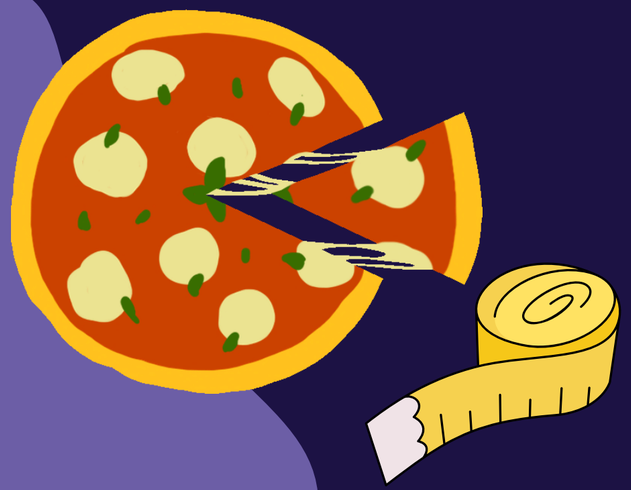
Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         612               96.6825 %
Incorrectly Classified Instances        21                3.3175 %
Kappa statistic                          0.9612
Mean absolute error                      0.0163
Root mean squared error                  0.0957
Relative absolute error                  6.6564 %
Root relative squared error             27.371  %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,011    0,939      1,000    0,968      0,964    0,994     0,939     Normal_Weight
                 0,927    0,004    0,974      0,927    0,950      0,943    0,987     0,932     Overweight_Level_I
                 0,964    0,007    0,952      0,964    0,958      0,952    0,992     0,934     Overweight_Level_II
                 0,982    0,011    0,948      0,982    0,965      0,957    0,984     0,934     Obesity_Type_I
                 0,961    0,000    1,000      0,961    0,980      0,978    0,997     0,978     Insufficient_Weight
                 0,934    0,002    0,988      0,934    0,960      0,955    0,990     0,951     Obesity_Type_II
                 0,990    0,004    0,980      0,990    0,985      0,982    0,993     0,971     Obesity_Type_III
Weighted Avg.    0,967    0,006    0,968      0,967    0,967      0,961    0,991     0,948

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  92   0   0   0   0   0   0 |   a = Normal_Weight
   3  76   3   0   0   0   0 |   b = Overweight_Level_I
   0   2  80   1   0   0   0 |   c = Overweight_Level_II
   0   0   1 109   0   0   1 |   d = Obesity_Type_I
   3   0   0   0  74   0   0 |   e = Insufficient_Weight
   0   0   0   5   0  85   1 |   f = Obesity_Type_II
   0   0   0   0   0   1  96 |   g = Obesity_Type_III
```

Figura 8 - Resultado do modelo J48 Tree com CfsSubsetEval com IMC. Fonte: Autores.

Treino e teste

Modelo Random Forest com CfsSubsetEval com IMC.



Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances         607               95.8926 %
Incorrectly Classified Instances        26                4.1074 %
Kappa statistic                          0.952
Mean absolute error                      0.0125
Root mean squared error                  0.0858
Relative absolute error                  5.1076 %
Root relative squared error             24.5231 %
Total Number of Instances              633

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,967    0,009    0,947      0,967    0,957      0,950    0,997     0,971     Normal_Weight
                 0,927    0,009    0,938      0,927    0,933      0,923    0,980     0,963     Overweight_Level_I
                 0,928    0,007    0,951      0,928    0,939      0,930    0,998     0,979     Overweight_Level_II
                 0,955    0,006    0,972      0,955    0,964      0,956    1,000     0,998     Obesity_Type_I
                 0,974    0,005    0,962      0,974    0,968      0,963    1,000     0,999     Insufficient_Weight
                 0,967    0,009    0,946      0,967    0,957      0,949    0,999     0,997     Obesity_Type_II
                 0,990    0,002    0,990      0,990    0,990      0,988    1,000     0,999     Obesity_Type_III
Weighted Avg.    0,959    0,007    0,959      0,959    0,959      0,952    0,997     0,987

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  89   0   0   0   3   0   0 |   a = Normal_Weight
   3  76   3   0   0   0   0 |   b = Overweight_Level_I
   0   5  77   1   0   0   0 |   c = Overweight_Level_II
   0   0   1 106   0   4   0 |   d = Obesity_Type_I
   2   0   0   0  75   0   0 |   e = Insufficient_Weight
   0   0   0   2   0  88   1 |   f = Obesity_Type_II
   0   0   0   0   0   1  96 |   g = Obesity_Type_III
```

Figura 9 - Resultado do Random Forest com CfsSubsetEval com IMC. Fonte: Autores.

Obrigado!

Dúvidas?

Contatos:
Isabela — isabelaloebel@gmail.com
Nickolas — nick.cremaa@gmail.com