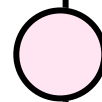
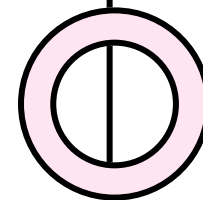


A P R E N D I Z A G E M A U T O M Á T I C A

PAULO REIS 1081376

PEDRO ALLEN 1211266

RITA AZEVEDO 1231439



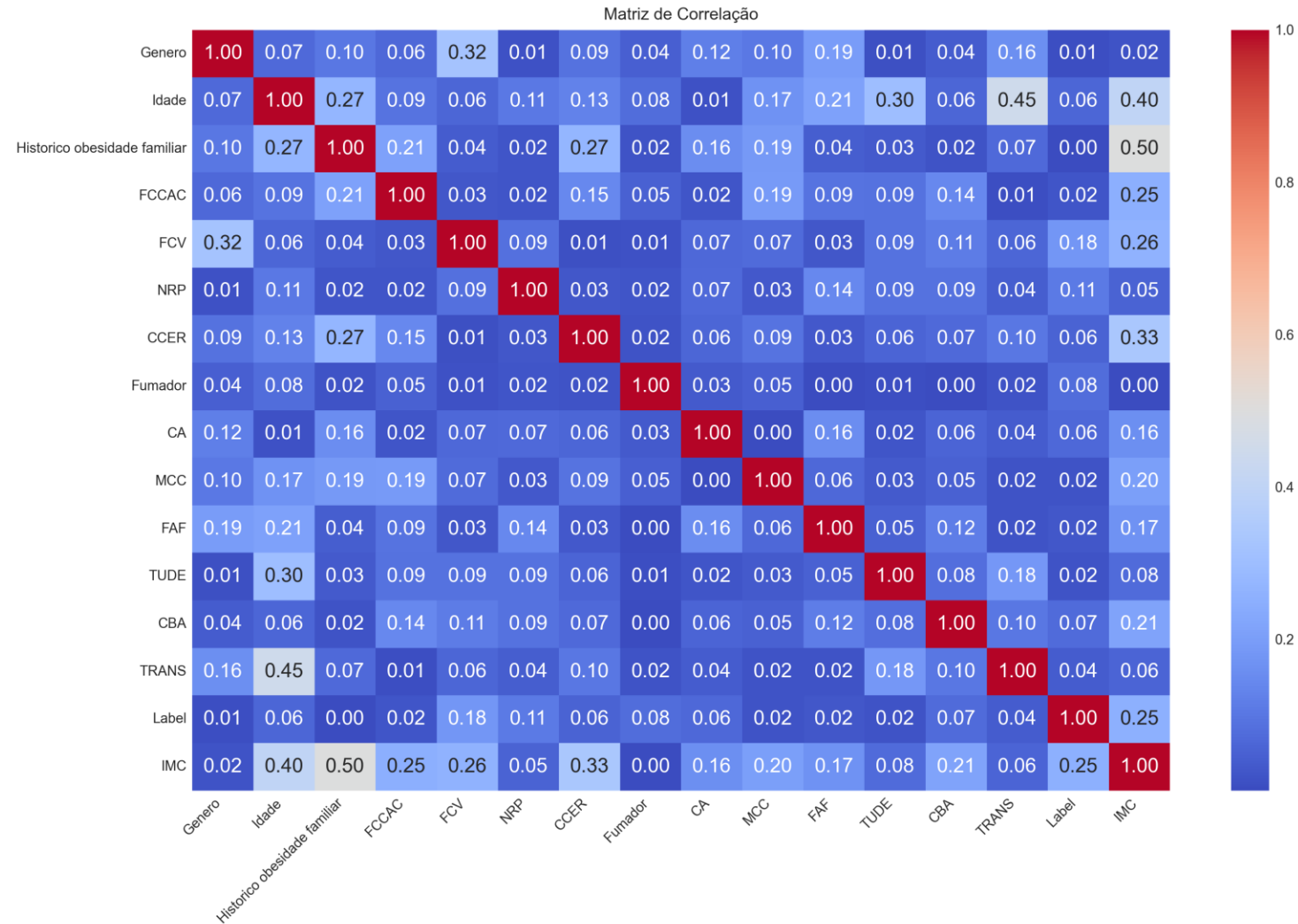
○ Exploração dos dados

- Estudamos cada uma das colunas para perceber que relação tinham.
- Analisamos dados relacionados ao longo dos anos, verificando-se diversas tendências e padrões.
- A distribuição do IMC apresentou uma distribuição normal.






Matriz de Correlação



○ Preparação dos dados

- Não se verificou existência de NaN, por isso não houve necessário remover valores nulos.
- Uso do LabelEncoder para transformar as classes em valores numéricos.
- Utilização do `pandas.get_dummies` para atributos com múltiplas classes.





MODELOS DE REGRESSÃO

○ Regressão Linear Simples

- Todos os preditores são independentes uns dos outros.
- De todos os preditores o histórico de obesidade familiar é o que apresenta menor erro na previsão.

Atributo	MAE	RMSE
Idade	6.38	7.77
Genero	6.58	7.98
HOF	5.73	7.075
FCCAC	6.41	7.73
FCV	6.40	7.74
NRP	6.58	8.00
CCER	6.01	7.39
Fumador	6.60	8.01
CA	6.55	7.97
MCC	6.39	7.80
FAF	6.49	7.87
TUDE	6.56	7.96
CBA	6.53	7.77
TRANS	6.60	7.98



○ Regressão Linear Múltipla

- Os erros foram menores quando aumentamos o número de preditores.
- Obtém-se uma melhor previsão.

R squared	47.835
MAE	4.5282
MSE	31.935
RMSE	5.6511



○ Árvore de decisão

- Apresenta melhores resultados que a Regressão Linear Múltipla.
- Requer mais afinação.

MAE	2.911
RMSE	3.496



○ MLPRegressor

- O complexo de afinação, tem muitos parâmetros.
- Os resultados variam a cada execução, para a mesma configuração.





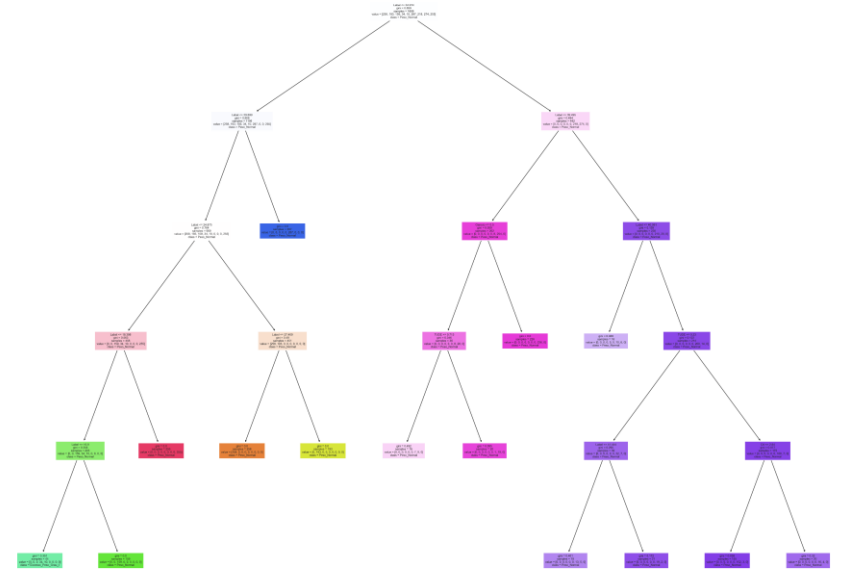
M O D E L O S D E C L A S S I F I C A Ç Ã O





Árvore de decisão

- Profundidade 5 e 16 amostras mínimas.
- Obteve-se uma acurácia de 98%.
- Obteve-se uma avaliação k-fold e observamos que os valores para cada fold variam entre 0.90 e 1. Uma acurácia media de 0,95 com desvio padrão de 0,049.

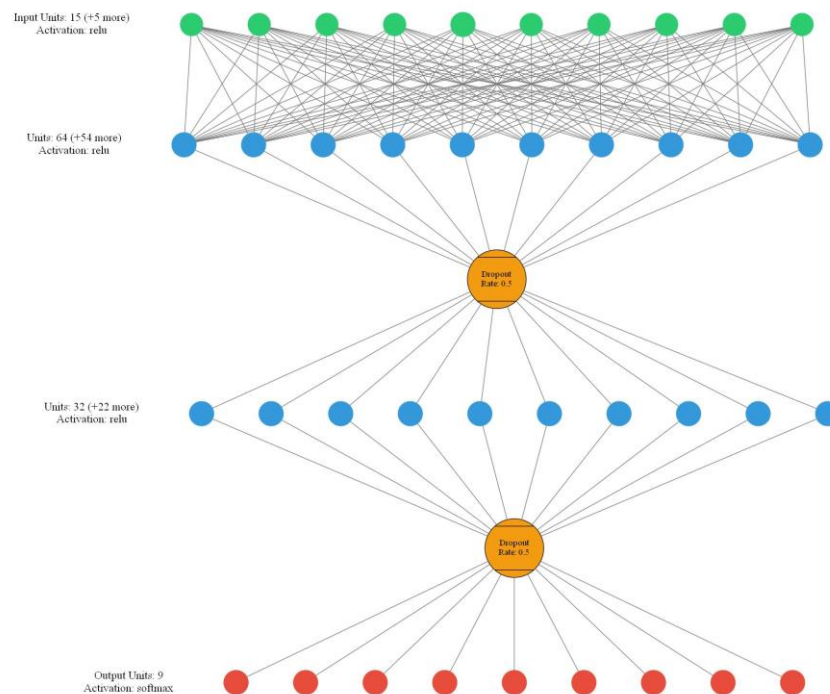


Árvore de decisão

- Obtemos um valor MSE de 22% e um valor RMSE de 2.43 prevendo a classe “Magreza Grau I”.

Classification Report:

	precision	recall	f1-score	support
0	0.64	0.98	0.77	50
1	0.95	0.73	0.82	51
2	0.51	0.85	0.63	39
3	0.00	0.00	0.00	19
4	0.00	0.00	0.00	3
5	1.00	0.83	0.91	78
6	0.80	0.98	0.88	57
7	0.81	0.78	0.79	64
8	0.82	0.60	0.69	62
accuracy			0.77	423
macro avg	0.61	0.64	0.61	423
weighted avg	0.77	0.77	0.76	423



○ K-Nearest Neighbors (k-NN)

- Utilizamos a distância Manhattan e definimos 50 vizinhos.
- Calculamos o MSE e acurácia para cada vizinho. Obtivemos valores entre 0,79 e 0.65 em ordem decrescente.
- O vizinho 1 apresentou melhor acurácia.
- Observou-se a matriz de confusão para o vizinho 1 e a classe “Excesso de Peso I” obteve a melhor acurácia.

```
[[56  4  0  0  0  2  0  0  7]
 [ 7 34  2  1  0  2  0  0  2]
 [ 1  1 43  2  0  0  0  0  3]
 [ 0  0  6  1  1  0  0  0  1]
 [ 0  0  0  1  0  0  0  0  0]
 [ 1  1  0  0  0 67  0  3  2]
 [ 0  0  0  0  0  0 38  6  0]
 [ 0  0  0  0  0  1  4 53  0]
 [11  3  5  1  1  5  0  0 44]]
```



○ Support Vector Machine

- Utilizamos grid search para otimizar os parâmetros, o parâmetro de regularização C e funções kernel para projetar os dados, com possíveis valores linear, polinomial, sigmoid e Radial Basis Function(RBF).
- Os melhores parâmetros encontrados foi regularização 1000, gamma 1, kernel linear.
- Obteve-se valores da acurácia e F1-score de 96%.



○ Observações

- Remove-mos as colunas 'FCCAC', 'TRANS', 'CA', 'MCC', 'Idade', 'CCER', 'NRP' e 'FCV' nos diferentes modelos, obtendo em todos os modelos um valor ligeiramente maior na acurácia.
- Ao remover-se a coluna IMC removeu-se nos modelos uma queda acentuada nos valores de acurácia.
- O modelo da árvore de decisão demonstrou o melhor desempenho, não estando muito longe o modelo SVM.
- O modelo da rede neuronal revelou pior desempenho, seguido pelo k-nearest neighbors que foi o que representou 2º pior desempenho.



○ Conclusão

- Com os modelos de regressão conclui-se que o melhor fator de previsão é o Histórico de obesidade Familiar.
- Com os modelos de classificação concluímos que o IMC foi o atributo mais influente nos modelos.

