**Relatório do desempenho de modelos para classificação do nível do MHI**

O relatório está divido em duas partes principais, a primeira contém as simulações referentes aos testes com apenas a base de dados sociodemográficos, e a segunda parte refere-se aos testes com a base sociodemográfica unida à base WHOQOL. Todos os resultados nas tabelas são retirados de um K-Fold com K=10, a menos que seja explicitado o contrário.

# **Desempenhos com a base de dados sociodemográficos:**

## **Desempenho sem manipulação de dados:**

No primeiro momento foi realizado um teste com vários modelos sem de balanceamento de dados, seleção de atributos e tunning. Os modelos têm os parâmetros padrões da biblioteca de origem, e o modelo **Dummy** não tem executa nenhum aprendizado, sua saída é sempre referente à classe com mais dominância.

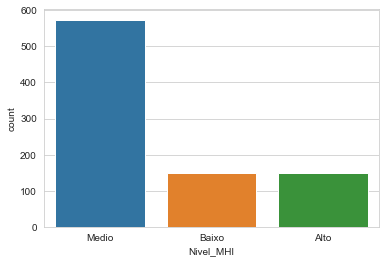
|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.650820 |
| RF | 0.652459 |
| SVM | 0.659016 |
| MLP | 0.573770 |
| XGB | 0.609836 |

A picture containing screenshot, colorfulness, rectangle, square

Description automatically generated

O resultado do **Dummy** é explicado pelo balancemanto dos dados.

|  |  |  |
| --- | --- | --- |
| **Classe** | **Total** | **Porcentagem** |
| Médio | 574 | 65.83 |
| Baixo | 149 | 17.09 |
| Alto | 149 | 17.09 |



## **Undersampling sem otimização**:

Assim, aplicando técnicas de Undersampling, nominalmente, Random Undersampling e NearMiss, e em seguida aplicando os testes obtemos os seguintes resultados.

### Random Undersampling:

|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.323129 |
| RF | 0.537415 |
| SVM | 0.302721 |
| MLP | 0.578231 |
| XGB | 0.564626 |

Gráfico, Gráfico de barras

Descrição gerada automaticamente

### NearMiss:

|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.323129 |
| RF | 0.384354 |
| SVM | 0.414966 |
| MLP | 0.479592 |
| XGB | 0.462585 |

Gráfico, Gráfico de barras

Descrição gerada automaticamente

## **Seleção de Atributos e Tunning:**

Assim, com esses resultados, percebe-se que a técnica Random Undersampling é a mais interessante para ser escolhida, juntamente com o modelo MLP os quais obtiveram, em média, o melhor resultado.

Logo, iremos aplicar as técnicas de seleção de atributos SFFS e SBFS nos dados balanceados, procurando a melhor combinação de atributos e utilizando como métrica de avaliação a acurácia do modelo. Assim, encontramos o seguinte resultado:

A picture containing line, plot, diagram, screenshot

Description automatically generated

Com os 10 melhores resultados de cada uma das técnicas sendo representados a seguir:

|  |  |
| --- | --- |
| **SFFS** | |
| **Num Atributos** | **Acurácia Média** |
| 33 | 0.703916 |
| 31 | 0.703916 |
| 32 | 0.70076 |
| 29 | 0.700643 |
| 27 | 0.693863 |
| 39 | 0.690532 |
| 28 | 0.690473 |
| 34 | 0.690415 |
| 35 | 0.687259 |
| 30 | 0.687084 |

|  |  |
| --- | --- |
| **SBFS** | |
| **Num Atributos** | **Acurácia Média** |
| 19 | 0.714202 |
| 18 | 0.714202 |
| 15 | 0.714144 |
| 17 | 0.714144 |
| 16 | 0.714085 |
| 20 | 0.707306 |
| 23 | 0.704091 |
| 14 | 0.703974 |
| 26 | 0.700701 |
| 24 | 0.700643 |

Assim, o que apresentou melhor média foi o SBFS, com apenas 19 atributos. Em seguida, utilizando os dados selecionados, foi aplicada a técnica de tunning dos Hyper parâmetros do modelo MLP, com uma acurácia media **de 0.6802454**, utilizando um k-fold com k=5.

|  |  |
| --- | --- |
| **Kth-Fold** | **Acurácia** |
| 1 | 0.593220 |
| 2 | 0.762712 |
| 3 | 0.728814 |
| 4 | 0.644068 |
| 5 | 0.672414 |

# **Desempenho com base de dados sociodemográfica e qualidade de vida (WHOQOL)**

## **Desempenho sem manipulação de dados:**

Assim, como base anterior, foi aplicado o mesmo procedimento, no entanto dessa vez é possível notar de imediato uma diferença considerável entre o modelo *Dummy* e o *RandomForest (RF)*.

|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.647541 |
| RF | 0.773770 |
| SVM | 0.740984 |
| MLP | 0.662295 |
| XGB | 0.773770 |

A picture containing screenshot, colorfulness, rectangle, square

Description automatically generated

O balanceamento dos dados inicial é o mesmo da base anterior.

|  |  |  |
| --- | --- | --- |
| **Classe** | **Total** | **Porcentagem** |
| Médio | 574 | 65.83 |
| Baixo | 149 | 17.09 |
| Alto | 149 | 17.09 |

A picture containing screenshot, text, rectangle, square

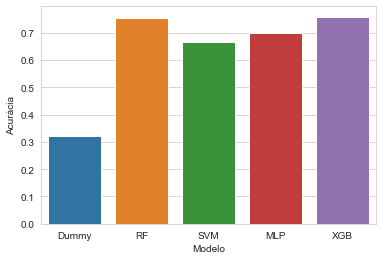
Description automatically generated

## **Undersampling sem otimização**:

Tratando-se do balanceamento de dados, novamente o *Random Undersampling* obteve um desempenho médio maior, além disso, nota-se que, diferentemente da base anterior, o modelo com melhor acurácia inicial não teve uma perda de acerto tão considerável.

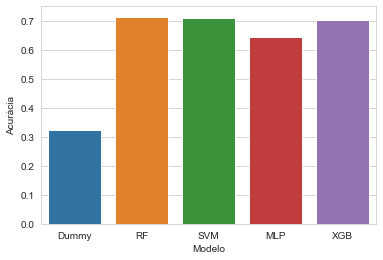
### Random Undersampling:

|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.323129 |
| RF | 0.755102 |
| SVM | 0.666667 |
| MLP | 0.697279 |
| XGB | 0.758503 |



### Near Miss:

|  |  |
| --- | --- |
| **MODELO** | **ACURACIA** |
| Dummy | 0.323129 |
| RF | 0.714286 |
| SVM | 0.710884 |
| MLP | 0.642857 |
| XGB | 0.704082 |



## **Seleção de Atributos e Tunning:**

As técnicas de seleção de atributos são as mesmas utilizadas anteriormente, com seus resultados expostos a seguir.

A screenshot of a graph

Description automatically generated with low confidence

|  |  |
| --- | --- |
| **SFFS** | |
| **Num Atributos** | **Acurácia Média** |
| 14 | 0.884278 |
| 18 | 0.880947 |
| 17 | 0.880888 |
| 15 | 0.88083 |
| 29 | 0.877557 |
| 30 | 0.877499 |
| 16 | 0.877499 |
| 13 | 0.877499 |
| 19 | 0.874167 |
| 26 | 0.874109 |

Assim, como o SFFS apresentou em média os 10 melhores resultados, ele foi escolhido para que fosse realizado o tunning dos hyper parâmetros, utilizando a melhor combinação de features e o modelo *RandomForest*, obtendo uma acurácia media de ***0.8672413****.* Uma tabela com o k-fold é apresentada a seguir.

|  |  |
| --- | --- |
| **Kth-Fold** | **Acurácia** |
| 1 | 0.900000 |
| 2 | 0.900000 |
| 3 | 0.866667 |
| 4 | 0.833333 |
| 5 | 0.827586 |
| 6 | 0.862069 |
| 7 | 0.862069 |
| 8 | 0.931034 |
| 9 | 0.827586 |
| 10 | 0.862069 |

|  |  |
| --- | --- |
| **SBFS** | |
| **Num Atributos** | **Acurácia Média** |
| 23 | 0.877499 |
| 13 | 0.87744 |
| 9 | 0.877382 |
| 26 | 0.874167 |
| 25 | 0.87405 |
| 24 | 0.870777 |
| 37 | 0.870777 |
| 36 | 0.87066 |
| 48 | 0.87066 |
| 49 | 0.867387 |