Relatório do desempenho de modelos para classificação do nível do MHI

O relatório está estruturado para de forma que forma que os resultados são apresentados de acordo com cada etapa de aprimoramento do modelo para cada uma das bases de dados utilizadas. Os testes foram feitos para as seguintes bases:

**1.** Somente o sociodemográfico;

**2.** Sociodemográfico e todas as respostas do WHOQOL; **3.** Sociodemográfico e os domínios do WHOQOL;

**4.** Sociodemográfico e o índice de qualidade de vida geral (Média das duas primeiras questões do WHOQOL);

**5.** Sociodemográfico e a média dos domínios do WHOQOL.

O objetivo é classificar os níveis do inventário de saúde mental, que são: baixo, médio e alto. A base de dados original apresentava um desbalanceamento, que pode ser visto na figura abaixo. Para tratar o desbalanceamento foi optado por utilizar uma técnica de *Undersampling*, nominalmente, *Random UnderSampling*.

Todos os resultados de desempenho nas tabelas são retirados de um *K-Fold* com K=5, a menos que seja explicitado o contrário.

1 DESEMPENHO SOMENTE COM BALANCEAMENTO DE DADOS:

A seguir é apresentado o desempenho de cada um dos modelos, tendo como única alteração o balanceamento de dados através da técnica de Undersampling. Com os resultados é possível perceber que com um modelo sem nenhum tipo de *tunning* o Random Forest tem um melhor desempenho em todas as bases excluindo o caso "Sociodemográfico e média dos domínios do WHOQOL", onde o *XGboost* apresenta por volta de 4% de melhoria.

1.1 SOMENTE O SOCIODEMOGRÁFICO;

**Modelo** **Acurácia** Dummy 0.316527 RF **0.481793** SVM 0.375350 MLP 0.417367 XGB 0.448179

**1.2** SOCIODEMOGRÁFICO E TODAS AS RESPOSTAS DO WHOQOL

**Modelo** **Acurácia** Dummy 0.313725 RF **0.742297** SVM 0.714286 MLP 0.616246 XGB 0.719888

1.3 SOCIODEMOGRÁFICO E OS DOMÍNIOS DO WHOQOL;

**Modelo** **Acurácia** Dummy 0.352941 RF **0.761905** SVM 0.675070 MLP 0.621849 XGB 0.711485

1.4 SOCIODEMOGRÁFICO E O ÍNDICE DE QUALIDADE DE VIDA GERAL;

**Modelo** **Acurácia** Dummy 0.277311 RF 0.565826 SVM 0.498599 MLP 0.481793 XGB **0.574230**

1.5 SOCIODEMOGRÁFICO E MÉDIA DOS DOMÍNIOS DO WHOQOL.

**Modelo** **Acurácia** Dummy 0.246499 RF **0.672269** SVM 0.509804 MLP 0.546218 XGB 0.680672

2 GRIDSEARCH COM BASES BALANCEADAS

Com os resultados dos modelos obtidos, foi realizado um *GridSearch* para poder *tunnar* os Hyper parâmetros dos modelos. Com o *tunning* feito os modelos com melhor desempenho em cada uma das bases foi selecionado para enfim ser realizado a escolha dos atributos. É importante ressaltar a melhora de desempenho encontrada com a base “Sociodemográfico e o índice de QV”, assim como o baixíssimo ganho de desempenho para a base “Sociodemográfico”.

Na tabela a seguir é possível visualizar os desempenhos dos modelos após o *tunning*, os modelos com a acurácia destacada em negrito foram usados para o passo de seleção de atributos.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Modelos** | | | | |
| **Base de Dados** | | **RF** **SVM** **MLP** **XGBoost** | | | |  |
| Sociodemográfico | | **0.668571** | 0.651429 | 0.662857 | 0.651429 | |
| Sociodemográfico e todas as respostas do  WHOQOL  Sociodemográfico e os domínios do WHOQOL; |  | **0.744444**    **0.771429** | 0.733333 | 0.688889 | 0.711111 | |
|  | 0.742857 | 0.731429 | 0.742857 | |
| Sociodemográfico e o índice de QV | | 0.674286 | **0.702857** | 0.702857 | 0.685714 | |
| Sociodemográfico e média dos domínios do  WHOQOL | | 0.680000 | 0.725714 | 0.714286 | **0.742857** | |

3 SELEÇÃO DE ATRIBUTOS

Finalmente, realizando a seleção de atributos para cada um dos modelos escolhidos para cada uma das bases, sendo eles:

• Sociodemográfico – Randon Forest – (Data 1)

• Socdem. e todas as respostas do WHOQOL – Random Forest – (Data 2) • Socdem. e domínios do WHOQOL – Random Forest – (Data 3)

• Socdem. e índices de QV – Support Vector Machine – (Data 4) • Socdem. e média dos domínios WHOQOL – XGBoost – (Data 5)

Para realizar a seleção de atributos foi utilizada as duas técnicas recomendadas: SBFS e SFFS. Com isso é possível observar os resultados com cada uma das técnicas. A legenda representa a ordem apresentada no início do documento.

**SFFS**

**Base** **Acurácia** Data 1 0.700670 Data 2 0.798207 Data 3 **0.798213** Data 4 0.700690

Data 5 0.770680

**SBFS**

**Base** **Acurácia** Data 1 0.702949 Data 2 **0.801668** Data 3 0.793635 Data 4 0.698404

Data 5 0.772959

É possível observar que ambos os métodos obtiveram resultados praticamente idênticos, com diferenças de menos de 1% na acurácia. Além disso, notamos uma melhoria considerável no desempenho para base contendo apenas os dados sociodemográficos.

Analisando a matriz de confusão para os melhores casos de cada método, identificamos uma forte tendência dos modelos em confundir o nível "Médio" com o nível "Alto".

Matriz de Confusão – SBFS – DATA 2 Matriz de Confusão – SFFS – DATA 3

Com base nessas informações, podemos concluir que os modelos que apresentaram melhor desempenho foram "Sociodemográfico com todas as respostas do WHOQOL" e "Sociodemográfico e os domínios do WHOQOL". Ao comparar os atributos selecionados pelo método SBFS, que totalizam quatorze e oito para Data 2 e 3, respectivamente, notamos que apenas uma delas é comum entre as bases mencionadas, que é o "Domínio Psicológico".

Por outro lado, ao analisarmos o método SFFS, onde são selecionados dezessete e dezesseis atributos para Data 2 e 3, respectivamente, percebemos que cinco atributos foram escolhidos para ambas as bases: "Curso é Estressante", "Atividade de Lazer", "Domínio Psicológico", "Turno do Curso" e "Situação Civil".

Os atributos selecionados pelo método SBFS para as duas bases mencionadas são listados a seguir:

|  |  |
| --- | --- |
| **Atributos para Data 2:** • Atividades profissionais:  • Atividade de Lazer:  o Ficar com a família; o Outros;  • Raça: Parda  • Tipo de Transporte para universidade  • Como você avaliaria sua qualidade de vida?  • Quão satisfeito(a) você está com a sua saúde?  • O quanto você consegue se concentrar? • Você tem dinheiro suficiente para  satisfazer suas necessidades?  • Quão satisfeito(a) você está consigo mesmo?  • Com que frequência você tem sentimentos negativos ?  • DOMINIO 1(FISICO);  • DOMINIO 2(PSICOLOGICO). | **Atributos para Data 3:**  • Já reprovou alguma disciplina/ modulo? • Em termos gerais, você considera seu  curso estressante?  • Quanto a Bolsas ou Auxílio Universitário: • Atividade de Lazer:  o Assistir Filme/Série • Raça: Amarela  • Atualmente reside com?  • Qual a sua religião?  • DOMINIO 2(PSICOLOGICO). |

Esses resultados indicam que o "Domínio Psicológico" é um atributo relevante e consistente para ambos os métodos de seleção de atributos. Além disso, a análise do método SFFS revela que os atributos "Curso é Estressante", "Atividade de Lazer", "Turno do Curso" e "Situação Civil" também desempenham papéis importantes na predição do desempenho dos modelos.