A3 - Comitê de Classificadores

Inicialmente escolhemos os dados na kaggle: Mental Health Diagnosis and Treatment Monitoring e focamos em tentar prever qual o melhor tipo de terapia em razão a determinado problema.

Escolhemos os seguintes modelos de I.A:

- NaiveBayes - GaussianNB
- DecisionTree - DecisionTreeClassifier
- SVM - SVC
- KNN - KNeighborsClassifier
- NeuralNetwork - MPLClassifier

1 - Limpeza de dados irrelevantes

Primeiramente foi feita uma análise dos dados brutos baseada na descrição de cada campo na base do Kaggle. Inicialmente foram removidas as colunas “Patient ID”, “AI-Detected Emotional State” e “Treatment start date”, pois eram desnecessárias, não agregando valor e consumindo processamento.

2 - Análise dos dados brutos

Em seguida, criamos um DataFrame e utilizamos o método “describe” para entender o comportamento individual dos dados, bem como o método “corr” para analisar se existe correlação entre as variáveis. A partir dessa análise concluímos que as variáveis estão fracamente correlacionadas.

3 - Transformação dos dados

Antes de prosseguir com os testes, foi necessário transformar as colunas de texto, atribuindo um ID para cada texto distinto.

4 - Testes dos métodos de IA

Para os testes, os dados foram filtrados para treinar o modelo apenas com os dados onde os pacientes obtiveram melhora, sendo y = “TherapyTypeId” e X os campos restantes (utilizando o campo ID gerado para os campos de texto ao invés do texto em si).

Ao utilizarmos esses dados para o treino, percebemos que os modelos não obtiveram uma boa acurácia, conforme mostrado abaixo:

	Model	Accuracy
0	NaiveBayes	0.268908
1	DecisionTree	0.260504
2	SVM	0.285714
3	KNN	0.235294
4	NeuralNetwork	0.282913

Além disso, o treino estava demorando para processar devido aos recursos limitados do Google Colaboratory, mesmo com ciclos baixos de repetições. O principal motivo era o modelo SVM - SC, então para melhorar o tempo de processamento baixamos o código fonte e rodamos localmente no PC, melhorando consideravelmente o tempo de treino.

5 - Criação de variáveis derivadas

A fim de melhorar os resultados dos modelos, foram criadas variáveis derivadas a partir dos dados que temos até agora.

1. AgeGroupId: a partir da análise feita no tópico 2, sabemos que a menor idade é de 18 anos e a maior é de 60. Sendo assim, foram divididos 3 grupos com range de ~14 anos: 1 - 18 a 32 anos (exclusivo), 2 - 32 a 46 (exclusivo) e 3 - 46 anos ou mais.
2. [variável]ImprovedRate: foi criada uma função que cria uma chave composta “TherapyTypeId”/[variável], onde é calculado o percentual de melhora entre todos os pacientes com aquela combinação.
3. [variável]ByTherapyScore: foi criada uma função conforme feito em [variável]ImprovedRate, mas calculando um score, considerando a quantidade de dados e a taxa de melhoria. Segue a fórmula: [variável]ImprovedRate + log2(quantidade de pacientes com a chave composta “TherapyTypeId”/[variável]).

Com isso, foi classificado o “AgeGroupId” para cada paciente e foram calculados os campos [campo]ImprovedRate e [campo]ByTherapyScore para cada combinação de “TherapyTypeId” com as variáveis “DiagnosisId”, “AgeGroupId” e “MedicationId”. Após isso, foi feito merge individual de cada valor no DataFrame de pacientes a partir da chave composta respectiva (“TherapyTypeId”/[variável])

6 - Treino final dos métodos de IA

Para o treino final, foram alteradas as features, sendo X = “DiagnosisIdImprovedRate”, “DiagnosisIdByTherapyScore”, “AgeGroupIdImprovedRate”, “AgeGroupIdByTherapyScore”, “MedicationIdImprovedRate” e “MedicationIdByTherapyScore”. Utilizando esses dados para o treino, os modelos apresentaram uma melhoria significante, segue abaixo os resultados:

	Model	Accuracy
0	NaiveBayes	0.705882
1	DecisionTree	0.826331
2	SVM	0.689076
3	KNN	0.773109
4	NeuralNetwork	0.537815

7 - Conclusão

Embora os modelos iniciais tenham apresentado baixo desempenho, a aplicação de técnicas de transformação de dados e a criação de variáveis derivadas resultaram em uma melhoria considerável nos resultados, evidenciando que a qualidade dos dados e a representatividade das variáveis têm um impacto mais significativo do que a escolha do modelo em si. Esse aprendizado destaca a importância de investir na preparação e no pré-processamento adequado dos dados. Além disso, o projeto ressaltou a necessidade de uma infraestrutura robusta para suportar modelos computacionalmente mais intensivos, garantindo eficiência e escalabilidade.

Dentro desse contexto, a Árvore de Decisão se mostrou a melhor alternativa, possuindo uma acurácia consideravelmente acima dentre os outros modelos, boa eficiência computacional e facilidade de

interpretação (pode ser visualizada com o método plot_tree do scikit-learn).

Equipe:

Augusto de Souza Santos
RA: 722314266

Davi Gramm Bauer
RA: 122320356

Aline de Resende Barbosa
RA: 32225080

Victor Molinas Ribeiro Freire
RA: 824155626

Luiz Felipe Dias Bertochi
RA: 12624216458

Vinicius Sanches Weber Brandão
RA: 1272226279