MONITORAMENTO E ANÁLISE PREDITIVA DO ESTRESSE EM ESTUDANTES UNIVERSITÁRIOS

INF 493

TRABALHO FINAL

Ana Clara Guerra - 108205
Thales Barcelos - 116229

INTRODUÇÃO AO DATASET

Dataset: Student Stress Monitoring Datasets
Tema: Fatores que influenciam o estresse em estudantes universitários
Origem: Coletado via Google Forms — estudantes de 18–21 anos
Formato: CSV | 1100 instâncias | 22 variáveis (21 features + variável alvo)

Objetivo: Investigar como fatores psicológicos, fisiológicos, sociais, ambientais e acadêmicos se relacionam com o nível de estresse em estudantes

Tipos de dados: Numéricos discretos
Variável target: *stress_level*

- 0: nenhum estresse
- 1: Estresse positivo
- 2: Estresse negativo (angústia)

VARIÁVEIS

0 – 5

headache
sleep_quality
breathing_problem
noise_level
living_conditions
safety
basic_needs
peer_pressure
extracurricular_activities
bullying
academic_performance
study_load
teacher_student_relationship
future_career_concerns

0 – 21

anxiety_level

0 – 27

depression

0 – 3

blood_pressure (1 - 3)
social_support

0 – 30

self_esteem

0 – 1

mental_health_history

- Fisiológico
- Social
- Acadêmico
- Psicológico
- Ambiental

PRÉ-PROCESSAMENTO (COLAB 0)

- Valores Ausentes: Nenhum
- Valores Duplicados: Nenhum
- Outliers: Nenhum (Visualização com boxplots)
- Identificação: Adicionado *id_pessoa*
- Balanceamento: *df['stress_level'].value_counts(normalize=True) * 100*
- Normalização: ***MinMaxScale***
- Salvo: ***StressLevelDataset_Scaled.csv***

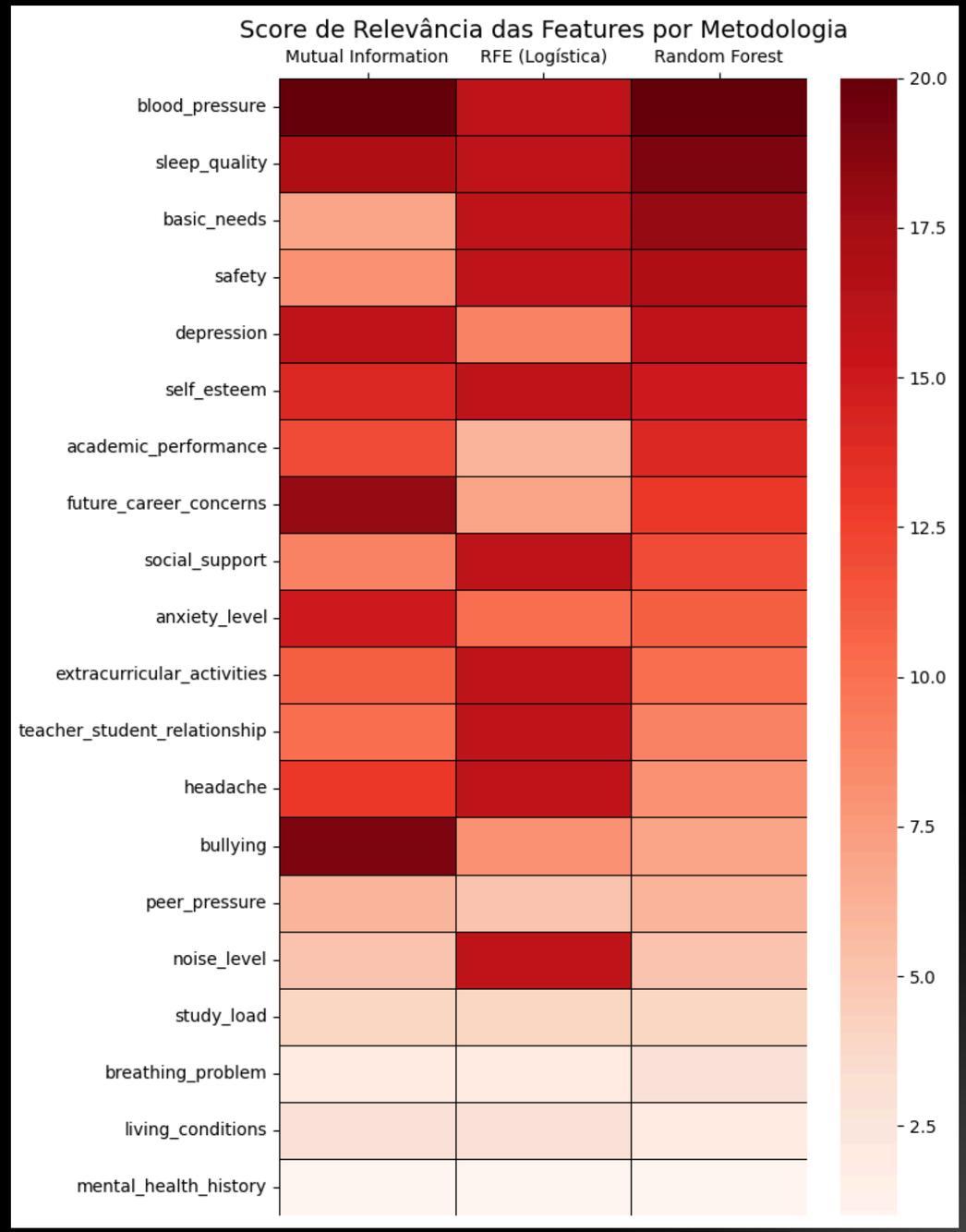
```
Em porcentagem (%):
stress_level
0    33.909091
2    33.545455
1    32.545455
```

id_pessoa	anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure	sleep_quality	breathing_problem	noise_level	...
1	0.666667	0.666667	0.0	0.407407	0.4	0.0	0.4	0.8	0.4	...
2	0.714286	0.266667	1.0	0.555556	1.0	1.0	0.2	0.8	0.6	...
3	0.571429	0.600000	1.0	0.518519	0.4	0.0	0.4	0.4	0.4	...
4	0.761905	0.400000	1.0	0.555556	0.8	1.0	0.2	0.6	0.8	...
5	0.761905	0.933333	0.0	0.259259	0.4	1.0	1.0	0.2	0.6	...

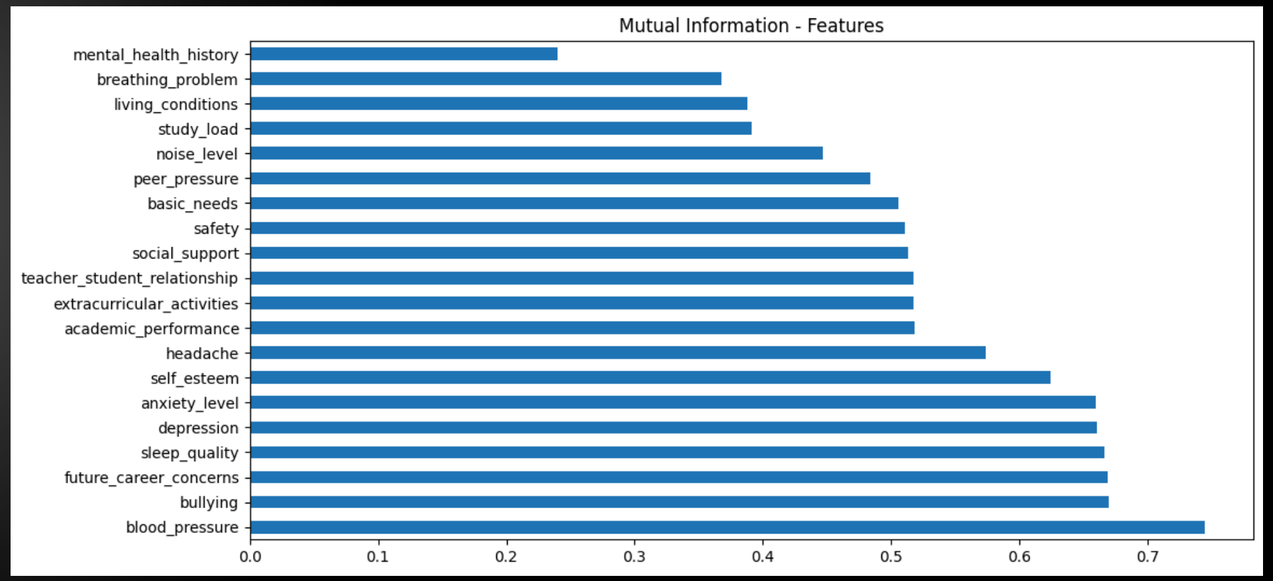
PREPARAÇÃO (COLAB 1)
SELEÇÃO DE VARIÁVEIS

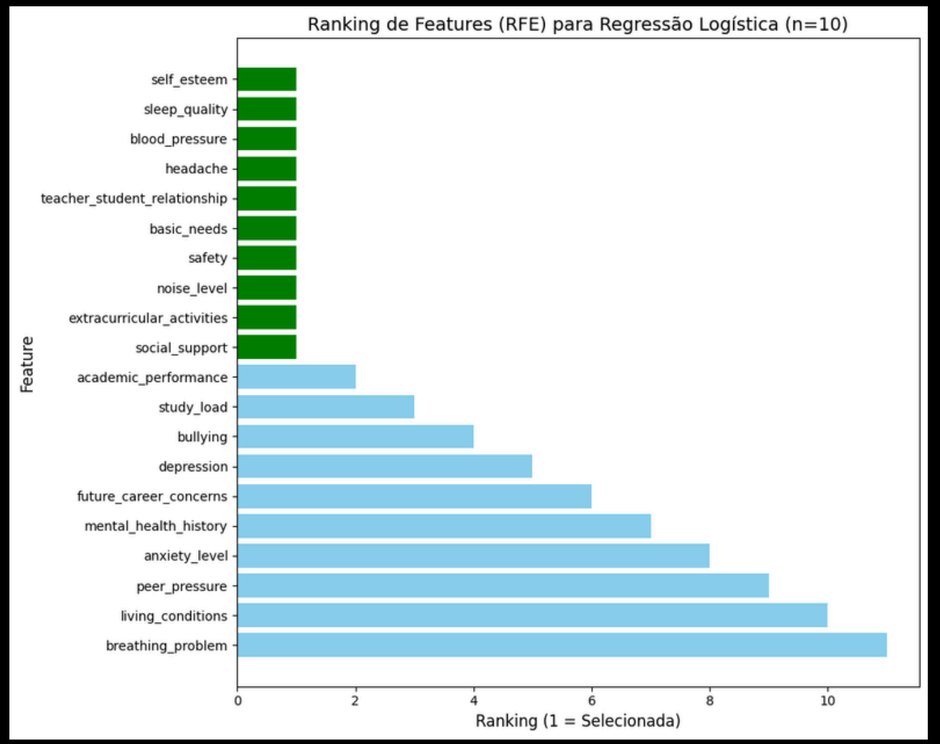
- Objetivo: Identificar e ranquear as variáveis mais relevantes
- Metodologias de Seleção de Atributos usadas
  - Filtro: *Mutual Information (MI)*
  - Wrapper: *Recursive Feature Elimination (RFE)* com Regressão Logística
  - Embedded: Importância pelo Random Forest (RF)

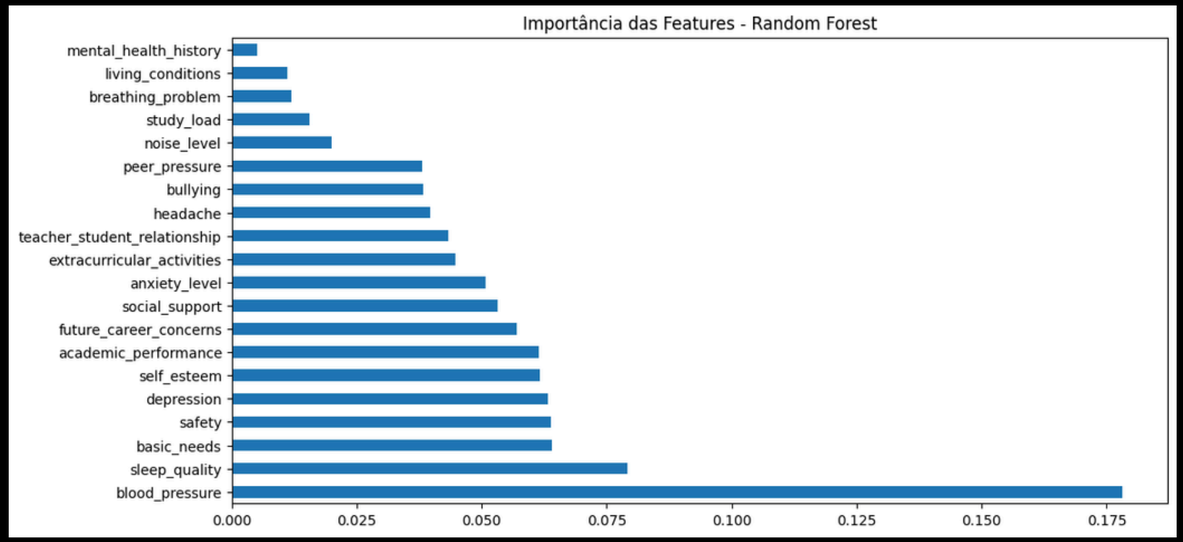
X_mi_top10.csv, X_mi_top5.csv, X_rfe_top10.csv, X_rfe_top5.csv, X_rf_top10.csv, X_rf_top5.csv e X_all.csv



PREPARAÇÃO (COLAB 1)

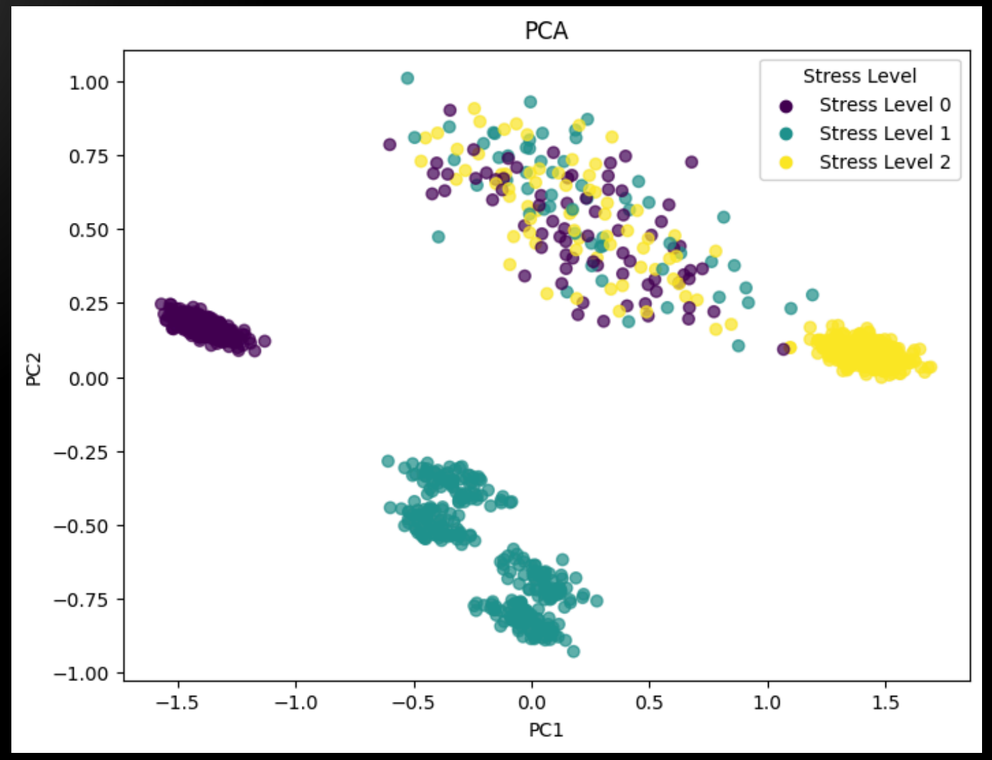


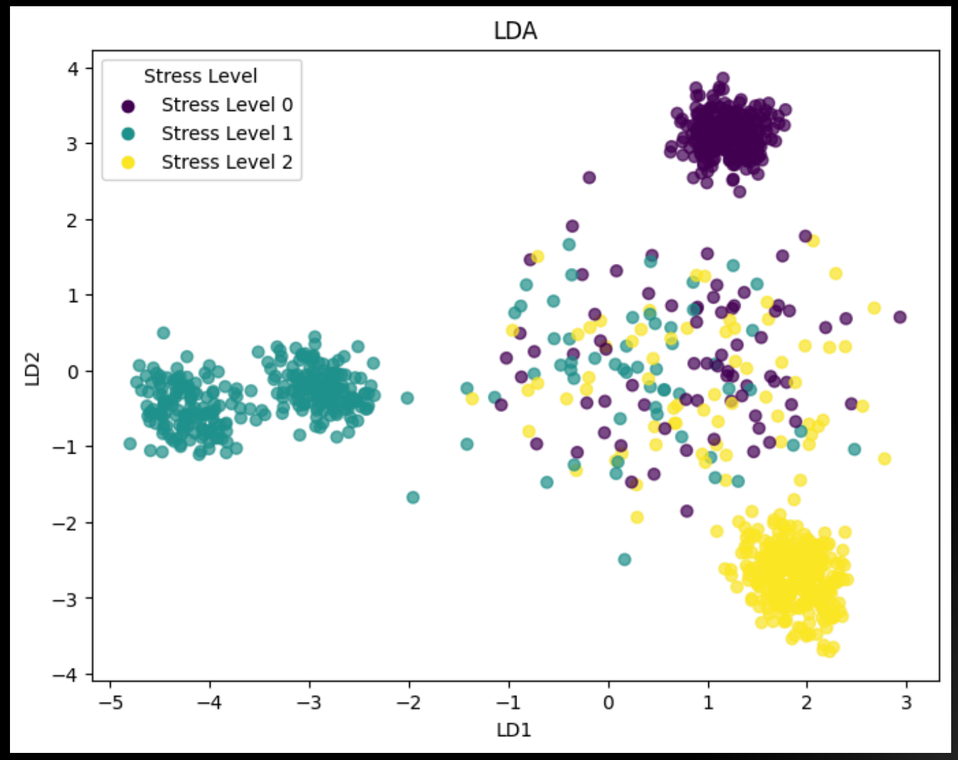




PREPARAÇÃO (COLAB 1)
REDUÇÃO DE DIMENSIONALIDADE

- Análise da possibilidade de reduzir a complexidade do dataset (20 features) para 2 dimensões, usando técnicas para visualizar a separabilidade das classes de estresse





MODELAGEM (COLAB 2)

ESTRATÉGIAS DE TESTE

- Divisão: Treino (80%) e Teste (20%) de forma estratificada
- Subconjuntos: 7 subconjuntos de features
  - Baseline: X_all.csv
  - Reduzidos: Top 10 e Top 5 das abordagens de Seleção de Atributos
- Métricas: Acurácia e F1 Score
- Algoritmos supervisionados de classificação
  - Regressão Logística (RL): Linear
  - KNN: Baseado em distância
  - Decision Tree (DT): Árvores
  - Random Forest (RF): Ensemble Bagging
  - SVM: Margem máxima
  - Gradient Boosting (GB): Ensemble Boosting
  - Multi-Layer Perceptron (MLP): Rede Neural Artificial

```
Treino: (880, 21)
Teste: (220, 21)
Distribuiçao treino:
stress_level
0    33.977273
2    33.522727
1    32.500000
Distribuiçao teste:
stress_level
0    33.636364
2    33.636364
1    32.727273
```

MODELAGEM (COLAB 2)
RESULTADOS

- ENTRAR NO COLAB

COMPARAÇÃO (COLAB 3)

- Comparação estatística para validar as diferenças de desempenho entre os modelos e escolher os melhores candidatos para a explicabilidade
- Metodologias
  - Agregação: Agrupamento dos resultados (Acurácia e F1) por modelo, calculando a média de desempenho através de todos os subconjuntos de features testados
  - Teste de Friedman
    - H0: todos os modelos com mesmo desempenho estatisticamente significativo entre os modelos
    - se p <= 0.05: existe diferença significativa
    - Resultado: p-value = 0.005307. **Logo, há diferença significativa entre pelo menos dois modelos**

COMPARAÇÃO (COLAB 3)

- Teste Pós-Hoc de Nemenyi
  - Identificar quais pares de modelos diferem significativamente (confiança 95%)
  - p > 0.05: diferenças NÃO significativas
  - p <= 0.05: diferença estatisticamente significativa (um modelo é melhor na media)

	Decision Tree	Gradient Boosting	KNN	MLP	Random Forest	Regressão Logística	SVM
Decision Tree	1.000000	0.997914	1.000000	0.314698	0.716494	0.026853	0.999796
Gradient Boosting	0.997914	1.000000	0.997914	0.676979	0.956377	0.126427	0.999997
KNN	1.000000	0.997914	1.000000	0.314698	0.716494	0.026853	0.999796
MLP	0.314698	0.676979	0.314698	1.000000	0.996250	0.956377	0.552164
Random Forest	0.716494	0.956377	0.716494	0.996250	1.000000	0.676979	0.903577
Regressão Logística	0.026853	0.126427	0.026853	0.956377	0.676979	1.000000	0.078926
SVM	0.999796	0.999997	0.999796	0.552164	0.903577	0.078926	1.000000

COMPARAÇÃO (COLAB 3)

- Seleção dos melhores modelos
  - **Multi-Layer Perceptron (MLP), Mutual Info Top 10**
    - Melhor desempenho, porém é um modelo caixa-preta
  - **Random Forest (RF), RFE Top 5**
    - Estatisticamente equivalente ao MLP, mas com a vantagem de oferecer maior interpretabilidade

Subconjunto	Mutual Info (Top 10)
Modelo	MLP
Acurácia	0.909091
F1 Score (macro)	0.909137
Subconjunto_Abreviado	MI (10)

Subconjunto	RFE (Top 5)
Modelo	Random Forest
Acurácia	0.895455
F1 Score (macro)	0.895544
Subconjunto_Abreviado	RFE (5)

EXPLICABILIDADE (COLAB 4)

- O **SHAP** e o **LIME** são ferramentas de Explicabilidade Artificial, etapa final do trabalho, que ajudam a entender as decisões dos modelos
- Ambas respondem à pergunta central: **"Por que o modelo tomou essa decisão específica?"** ou **"Quais features foram importantes para esta previsão?"**
- SHAP (SHapley Additive Explanations)
  - "Qual é a importância e a direção (positiva/negativa) de cada feature para a previsão, baseada em justiça e atribuição de valor?"
- LIME (Local Interpretable Model-agnostic Explanations)
  - Fornece explicações locais, contrastantes e compreensíveis (por exemplo, "A pressão alta aumentou em 20% a chance de Estresse Nível 2"

CONCLUSÃO

- O trabalho alcançou seu objetivo principal: identificar os fatores mais impactantes no estresse de estudantes universitários e construir modelos de classificação robustos para sua predição
- Principais Aprendizados e Validações
  - Fatores Chave de Estresse
    - O ranking combinado de features (MI, RFE, RF) mostrou que a previsão do estresse é dominada por uma combinação de fatores:
      - Fisiológicos: *blood_pressure* e *sleep_quality*
      - Psicossociais: *self_esteem, depression e future_career_concerns*
      - A variável *mental_health_history* foi consistentemente a de menor relevância
- Impacto da Seleção de Features
  - A redução do conjunto de dados de 21 para 5 e 10 features melhorou a performance preditiva, evitando o ruído que prejudicava modelos sensíveis à dimensionalidade (KNN, SVM)
- Robustez e Interpretabilidade
  - As análises de SHAP e LIME confirmaram que os modelos estam utilizando as features esperadas para realizar as previsões, validando a integridade de todo o pipeline
-

CONCLUSÃO

Comparação RF x MLP - Explicabilidade

89,55%	0,90	90,91%	90,91%
Acurácia RF	F1-Score RF	Acurácia MLP	F1-Score MLP

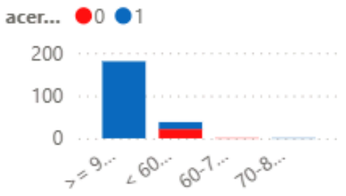
Matriz de Confusão RF

classe_real	0	1	2	Total
0	64	5	5	**74**
1	1	66	5	**72**
2	4	3	67	**74**
Total	**69**	**74**	**77**	**220**

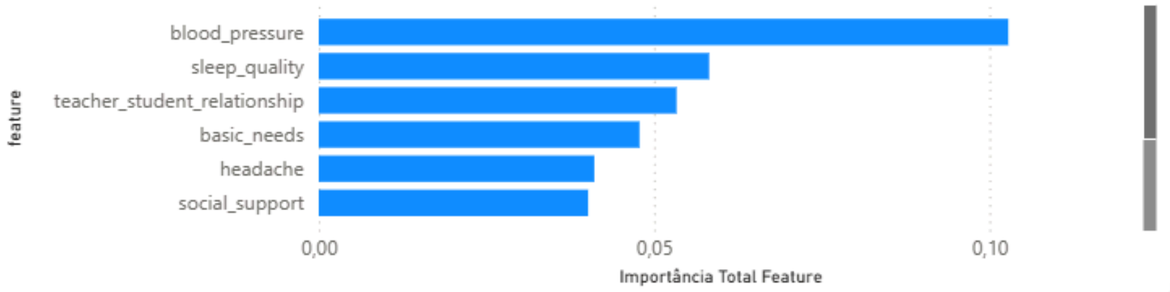
Matriz de Confusão MLP

classe_real	0	1	2	Total
0	67	2	5	**74**
1	1	68	3	**72**
2	3	6	65	**74**
Total	**71**	**76**	**73**	**220**

Total Predições por Faixa Confiança e acertou_rf



Importância Total Feature por feature



REFERÊNCIAS

- 0_PreProcessamento: https://colab.research.google.com/drive/1-dF7bfLRiwuSnwk_Emd-bNGI9DPKabiW
- 1_Preparacao (Seleção de Features): https://colab.research.google.com/drive/12TMq9QinrJk53nOUcdHPt6PHXV7PQw9u
- 2_Modelagem: https://colab.research.google.com/drive/1MrMSoSns5_hE-s_ojpmeuhjo0jOxCQw-?usp=sharing#scrollTo=9jrRipVzjLw
- 3_Comparacao (Análise Estatística): https://colab.research.google.com/drive/1P1fgxbkKofypOHrnoOe62Tcn1qSsYRtY#scrollTo=NFqSZ3lg2jmA
- 4_Explicabilidade (XAI): https://colab.research.google.com/drive/1v_uMYSRM9GRlBQLZ-ytp1H8nwK_GZ-X0#scrollTo=oAHrcSOuxc9O
- GitHub (Código e Dados): https://github.com/AnaClaraGuerra22/INF-493-CD/tree/4fc21c3e54b3aaddcd4f584e0460e097f4999fa9/TRAB%20FINAL
- Kaggle: https://www.kaggle.com/datasets/mdsultanulislamovi/student-stress-monitoring-datasets

OBRIGADO

DÚVIDAS?