Datathon do VI ENGOPE: equipe Embrapeiros

Carlos Eduardo Gonçalves de Oliveira, Luis Davi Araújo Pereira,
Paulo Augusto de Oliveira Gonçalves e Felipe Waks Andrade

17/10/2024

1 Motivação

A motivação do que preparamos para o Datathon do VI ENGOPE foi de atender ao objetivo geral de propor uma classificação de níveis de degradação (mostrando também um mapa temático para as mesmas) para as pastagens localizadas no estado de Goiás assim como de fornecer uma plataforma interativa com o qual o público possa analisar e explorar os dados disponíveis.

Assim, desenvolvemos um aplicativo web usando o pacote *streamlit* do *Python* de modo a tornar possível a requisição de predições por parte dos usuários e a disponibilidade de vários gráficos interativos, cujas variáveis podem ser alteradas conforme o desejo do usuário.

As seções incluídas no aplicativo web foram:

- **Predição da Degradação de Pastagens**: inclui informações sobre o modelo de classificação (XGBoost Chen (2015)) treinado sobre os dados disponíveis assim como um formulário que permite ao usuário obter as predições do modelo de acordo com diferentes valores para os escores. Com o usuário clicando em "Predizer", ele terá acesso também aos valores SHAP Rozemberczki et al. (2022) correspondentes (juntamente com um texto explicativo) mostrando como cada escore contribuiu para o aumento ou diminuição da probabilidade para a classe de degradação da pastagem (vide Figura 1);
- **Mapa temático das predições do modelo**: inclui um mapa mostrando como cada ponto disponível no conjunto de dados é classificado pelo modelo de predição, com verde representando pastagens não degradadas e vermelho, pastagens degradadas (vide Figura 2);
- **Análise Descritiva das variáveis**: inclui formas de visualização (boxplot e scatterplot) que permitem analisar como cada escore (ou par de escores) se relaciona com as classes de degradação de pastagem (vide Figura 3 e 4);
- **Análise de Discordância entre Avaliadores**: inclui formas de visualização (boxplots principalmente) que permitem analisar quais escores apresentaram as maiores diferenças (diferença entre avaliação máxima e mínima) para os diferentes escores, destrinchando se há relação com as classes de degradação (vide Figura 5 e 6);

O aplicativo web pode ser acessado pelo seguinte site: https://datathonengopevi.streamlit.app/.

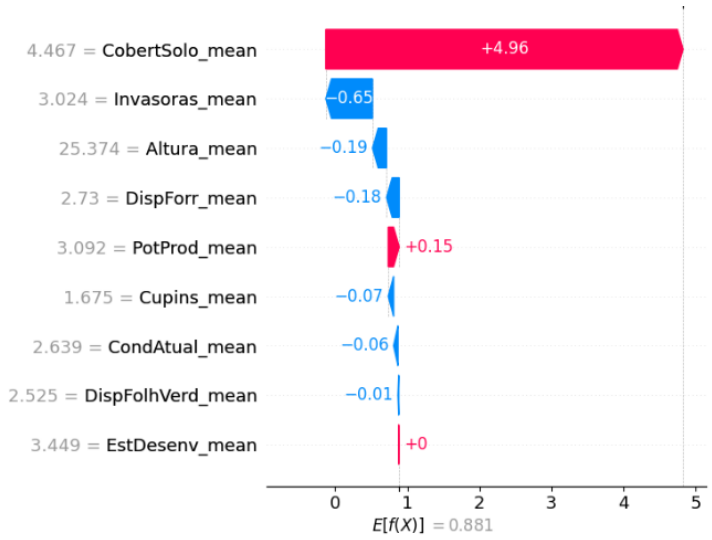

Figure 1: Figura resultante da explicação da contribuição das variáveis para uma predição qualquer utilizando o modelo de classificação disponível no aplicativo web.

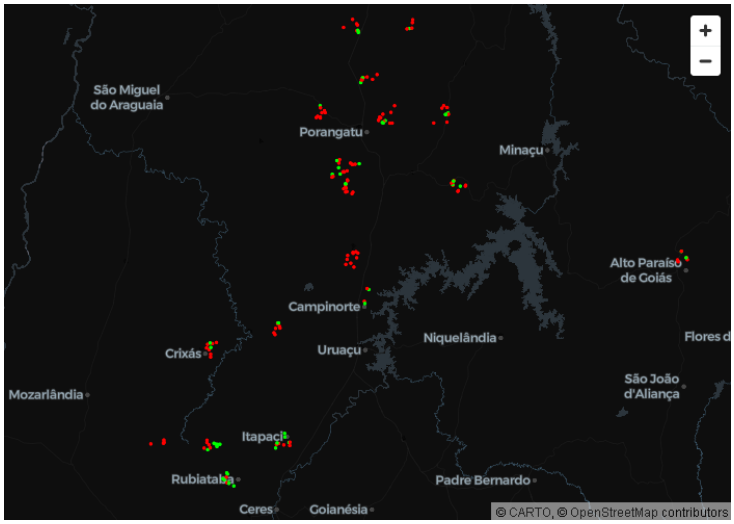

Figure 2: Print do mapa temático interativo incluído no aplicativo web resultante deste trabalho. Os valores em verde correspondem a pastagens não degradadas ou com pouca degradação, enquanto que a cor vermelha corresponde às pastagens degradadas.

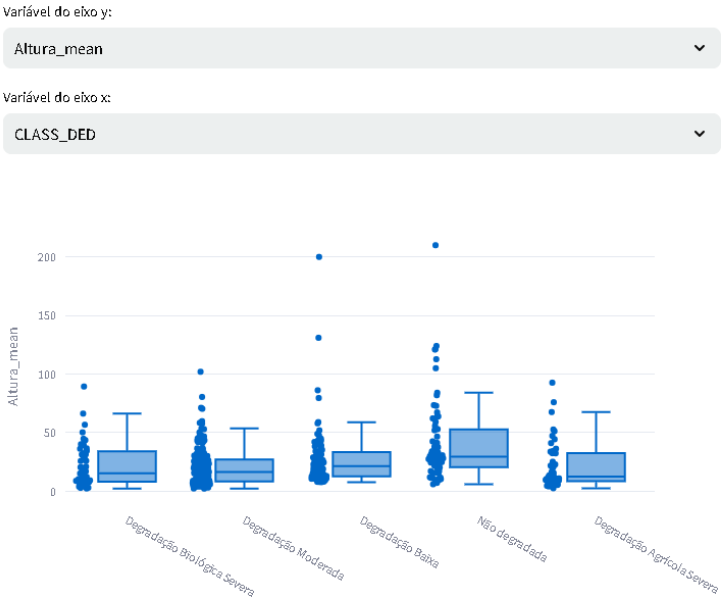

Figure 3: Exemplo de *boxplot* interativo utilizado para análise descritiva disponível no aplicativo web.

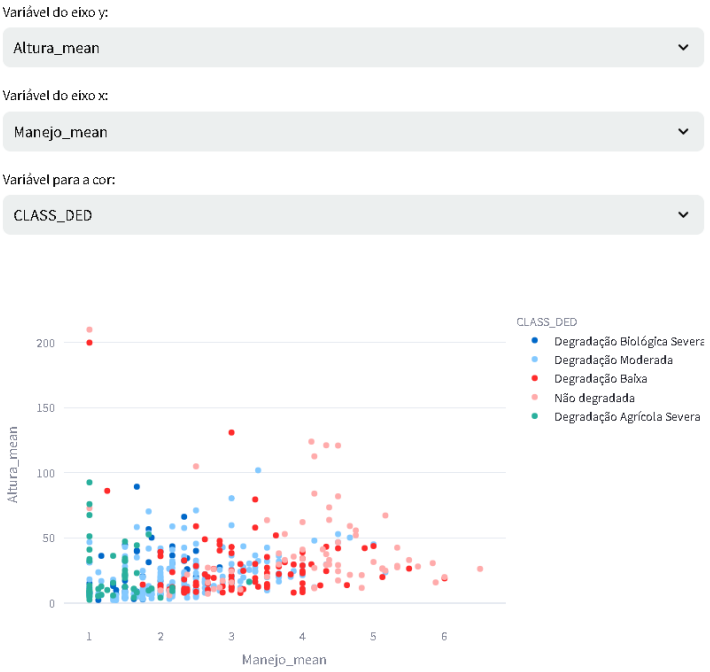

Figure 4: Exemplo de *scatterplot* interativo utilizado para análise descritiva disponível no aplicativo web.

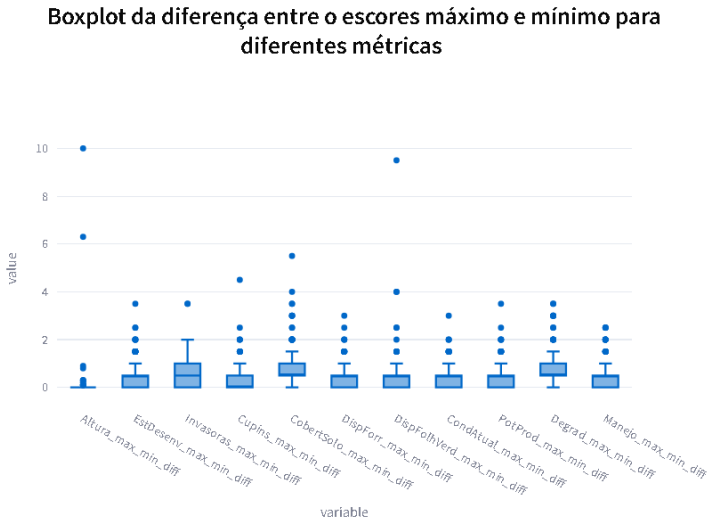

Figure 5: Exemplo de *boxplot* interativo utilizado para análise de discordância entre os avaliadores (usando a diferença entre o valor máximo e mínimo para um mesmo escore) para diferentes tipos de escore disponível no aplicativo web.

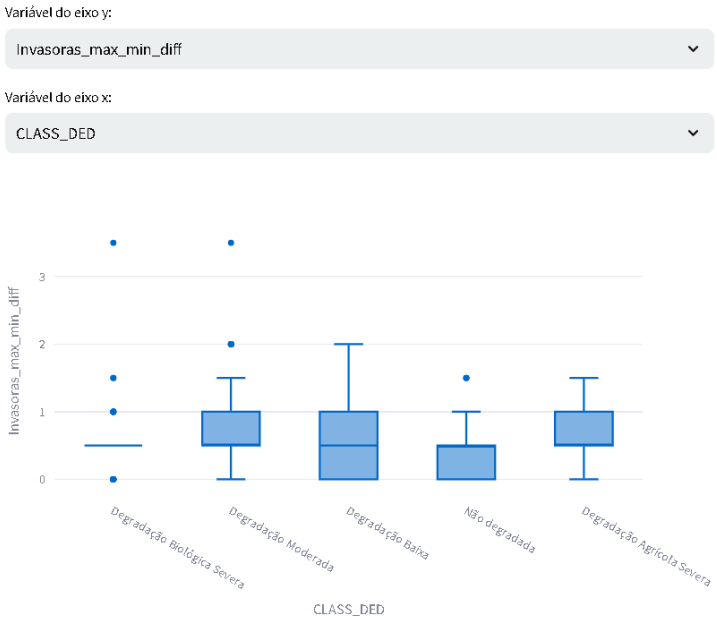

Figure 6: Exemplo de *boxplot* interativo utilizado para análise de discordância entre os avaliadores (usando a diferença entre o valor máximo e mínimo para um mesmo escore) avaliado para um mesmo escore e para diferentes classes de degradação de pastagem, disponível no aplicativo web.

2 *Storytelling*

2.1 Ponto de partida

Como dizia Lavoisier: "Nada se cria, nada se perde, tudo se transforma." Ao iniciar este projeto, partimos de uma classificação pré-existente dos tipos de pasto, que serviu como base inicial de análise. Avaliamos minuciosamente essa classificação inicial, identificando padrões e possíveis inconsistências. Nosso objetivo principal foi propor novas classificações que agrupassem os tipos de solo de maneira mais homogênea, facilitando a identificação de propriedades comuns. Utilizamos abordagens estatísticas e algoritmos de aprendizado de máquina para refinar os critérios de agrupamento e garantir que os novos grupos fossem mais representativos e consistentes com as características observadas. Essa transformação não só otimizou

a análise dos dados, como também forneceu uma base mais robusta para futuras tomadas de decisão em manejo agrícola.

2.2 Organização dos dados

Como há diferentes valores para os escores para diferentes pontos, foi necessário agrupá-los de alguma forma para cada ponto. Nesse caso, optamos por usar a média, de modo que cada ponto possuirá valores médios para os escores. Para uma análise inicial da classificação da degradação das pastagens, agrupamos as classes "Não degradada" e "Degradação Baixa" para compor a classe negativa "NÃO DEGRADADA"; enquanto que as classes "Degradação Agrícola Severa", "Degradação Biológica Severa" e "Degradação Moderada" foram separadas para compor, juntas, a classe "DEGRADADA". Fizemos isso pois, num primeiro momento, não entendemos quais variáveis mais influenciam na degradação das pastagens, de modo que uma análise considerando o caso binário é mais simples e pode fornecer conclusões relevantes. Vale ressaltar que não incluímos os escores "Degrad" e "Manejo" por darem uma informação direta demais acerca da degradação das pastagens.

2.3 XGBoost: modelo robusto para entender a relação entre as variáveis e a degradação das pastagens

Com os dados devidamente preparados e com as classes definidas ("NÃO DEGRADADA" e "DEGRADADA"), nós fizemos uso do modelo de classificação XGBoost para tentar reproduzir a classe binária. Como a interpretação das predições do XGBoost não é trivial, fizemos uso do framework SHAP para ajudar no entendimento da importância das variáveis preditoras para as predições.

Para uma validação robusta do modelo, separamos os dados em dados de treino (70% das instâncias: 322 pontos) e dados de teste (30% das instâncias: 138 pontos). Os dados de treino foram utilizados, para além de treinar o XGBoost, também para a tunagem de hiperparâmetros, que foi realizada usando-se *K-Fold Cross Validation* com 5 folds para determinar a melhor combinação de hiperparâmetros. O otimizador utilizado para busca no conjunto de hiperparâmetros foi o otimizador de Bayes. Após obter a melhor combinação para os hiperparâmetros, o modelo foi treinado sobre todos os dados de treino e em seguida aplicado sobre os dados de teste, alcançando a performance de, para a classe pastagem 'DEGRADADA': Precisão = 100%, Sensibilidade = 98.91%, F1-score = 99.45% e, para a classe pastagem 'NÃO DEGRADADA': Precisão = 97.87%, Sensibilidade = 100.00%, F1-score = 98.92%, resultando numa acurácia total de 99.28%.

Essa performance tão boa nos dados de teste nos deixou assustados, pois é incomum que sejam obtidos métricas tão boas e tão próximas de 100% para dados de teste. Nesse sentido, fizemos a análise usando o framework SHAP para averiguar que variáveis preditoras contribuíram para as predições, de modo a entendermos qual delas mais impactam na classe binária. O gráfico que obtivemos nesse sentido está mostrado na Figura 7, e indica contribuição extremamente alta para a variável "CobertSolo" e "Invasoras".

Figure 7: Valores SHAP resultantes das predições do modelo XGBoost treinado para classe binária de degradação das pastagens, indicando maior contribuição para as variáveis "CobertSolo" e "Invasoras".

2.4 "CobertSolo" e "Invasoras" servem bem para o problema multiclasse?

Tendo identificado essas duas variáveis como as mais importantes para o problema binário, vale a pena investigarmos a relação entre elas e as classes de referência provenientes de "CLASS_DED". Das Figuras 8, 9 e 10, vemos que "CobertSolo" e "Invasoras" servem muito bem para separar "CLASS_DED", de modo que isso sugere que talvez uma única árvore de decisão sirva para fazer uma boa classificação.

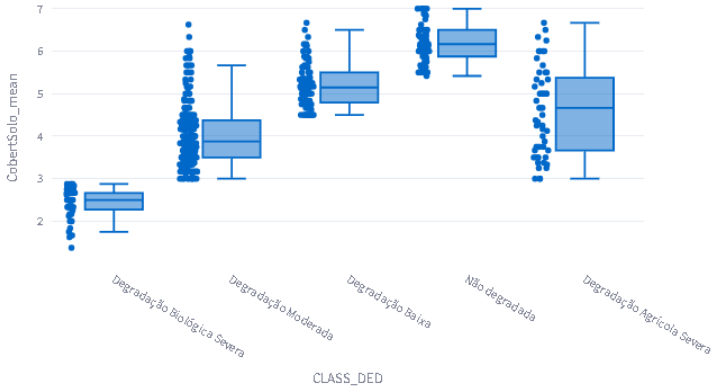

Figure 8: *Boxplot* mostrando os valores do escore "CobertSolo" de acordo com as diferentes classes de "CLASS_DED".

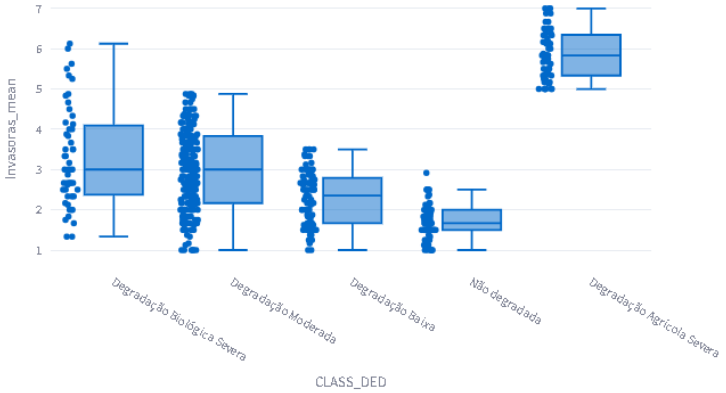

Figure 9: *Boxplot* mostrando os valores do escore "Invasoras" de acordo com as diferentes classes de "CLASS_DED".

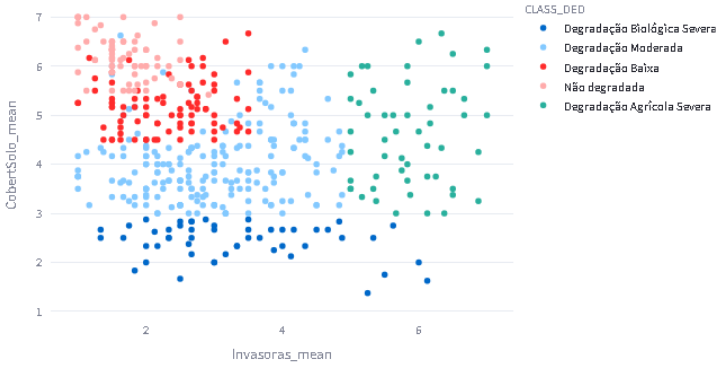

Figure 10: *Scatterplot* envolvendo "CobertSolo" e "Invasoras" coloridas por classe da "CLASS_DED". O *scatterplot* sugere que uma simples árvore de decisão pode ser capaz de separar as diferentes classes.

Conforme sugerido pela Figura 10, a classificação proveniente de "CLASS_DED" pode ser facilmente reproduzida por uma árvore de decisão envolvendo somente as variáveis "CobertSolo" e "Invasoras",

com a performance da árvore de decisão nos dados de teste correspondendo às métricas mostradas na Figura 11. Como árvores de decisão são facilmente interpretáveis, é possível plotar a regra de decisão das mesmas, e ela está mostrada na Figura 12.

	precision	recall	f1-score	support
0	0.7407	1.0000	0.8511	20
1	0.8462	0.8148	0.8302	27
2	1.0000	0.9000	0.9474	60
3	1.0000	1.0000	1.0000	16
4	1.0000	1.0000	1.0000	15
accuracy			0.9203	138
macro avg	0.9174	0.9430	0.9257	138
weighted avg	0.9323	0.9203	0.9223	138

Figure 11: Performance nos dados de teste de uma árvore de decisão treinada usando-se somente as variáveis "CobertSolo" e "Invasoras" para a predição das classes de "CLASS_DED".

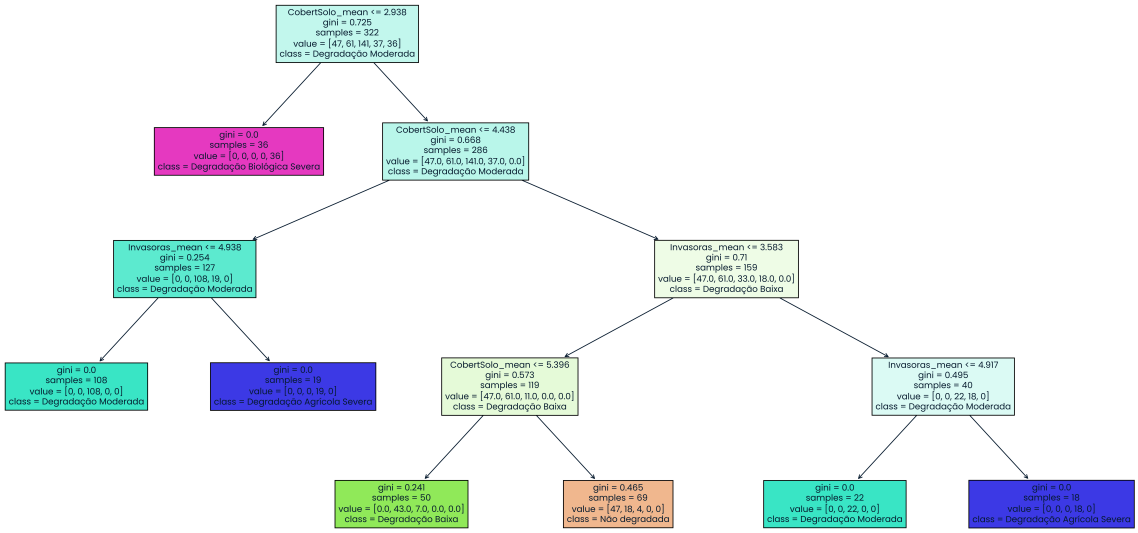

Figure 12: Árvore de decisão treinada usando-se somente as variáveis "CobertSolo" e "Invasoras" para a predição das classes de "CLASS_DED".

Para obter uma confirmação mais sólida sobre o impacto das variáveis "CobertSolo" e "Invasoras", o modelo XGBoost foi novamente empregado, por se tratar de um modelo ultra robusto e com capacidade de explorar de forma aprofundada as relações entre variáveis e covariáveis. Desta vez, em vez de utilizar o framework SHAP, optou-se pela função varImp no R, que se baseia nas informações de ganho de similaridade dos splits das árvores de decisão geradas pelo XGBoost para medir a importância das variáveis. Essa abordagem permite uma interpretação mais direta e eficiente da contribuição de cada variável no modelo.

xgbTree variable importance

	Overall
CobertSolo_media	100.000
Invasoras_media	72.347
Degrad_media	37.219
EstDesenv_media	5.359
Altura	3.079
Manejo_media	2.960
DispForr_media	2.821
DispFolhVerd_media	2.448
PotProd_media	2.338
Cupins_media	1.977
CondAtual_media	0.000

xgbTree variable importance

	Overall
CobertSolo_media	100.0000
Invasoras_media	65.5067
Degrad_media	54.8415
Cupins_media	15.0753
DispForr_media	13.1571
EstDesenv_media	9.7239
DispFolhVerd_media	4.4592
Altura	2.5880
Manejo_media	1.8326
CondAtual_media	0.2243
PotProd_media	0.0000

xgbTree variable importance

	Overall
CobertSolo_media	100.000
Invasoras_media	50.669
Degrad_media	9.495
Manejo_media	6.708
DispFolhVerd_media	4.239
Cupins_media	3.253
DispForr_media	3.206
EstDesenv_media	2.489
CondAtual_media	2.419
PotProd_media	2.194
Altura	0.000

xgbTree variable importance

	Overall
CobertSolo_media	100.000
Invasoras_media	62.216
Degrad_media	23.065
DispForr_media	14.474
EstDesenv_media	7.482
Cupins_media	4.968
Altura	3.431
Manejo_media	2.843
DispFolhVerd_media	1.241
CondAtual_media	1.052
PotProd_media	0.000

Figure 13: Valores varIMP resultantes do modelo XGBoost treinado para as classes propostas pela pesquisadora do LAPIG, indicando maior contribuição para as variáveis "CobertSolo" e "Invasoras".

Conforme ilustrado na 13, para cada *Fold* da validação cruzada, foram obtidos os respectivos valores de varImp. Observa-se que as variáveis "CobertSolo" e "Invasoras" destacam-se como predominantes na explicação das diferenças entre as classificações de degradação de solo propostas pelo LAPIG. Isso reforça a relevância dessas variáveis na modelagem preditiva e na compreensão dos padrões de degradação analisados.

3 Propondo uma nova classificação

A utilização de um método de classificação que pode ser facilmente descrito por apenas duas variáveis, embora possa trazer simplicidade e facilidade de interpretação, peca por não capturar a complexidade real dos dados. Esse tipo de abordagem limitada pode deixar de considerar importantes interações e influências de outras variáveis, resultando em uma visão superficial do fenômeno estudado e comprometendo a precisão das classificações. Ao restringir o número de variáveis, há também o risco de subutilizar informações valiosas que poderiam melhorar e enriquecer significativamente a característica das classes.

Propomos uma metodologia que segue os seguintes passos: primeiramente, definimos o número de clusters como 5, para manter a consistência com o número de níveis do método anterior. Em seguida, as covariáveis foram normalizadas, subtraindo suas respectivas médias e dividindo pelos seus erros padrão. Após a normalização, aplicamos uma Análise de Componentes Principais (PCA) Kurita (2019) aos dados, e selecionamos 5 componentes principais, pois juntas explicavam 90% da variância. Com os componentes selecionados, realizamos a clusterização utilizando o método K-means multidimensional Ahmed et al. (2020). Por fim, os clusters obtidos foram incorporados ao conjunto de dados, substituindo a classificação anterior.

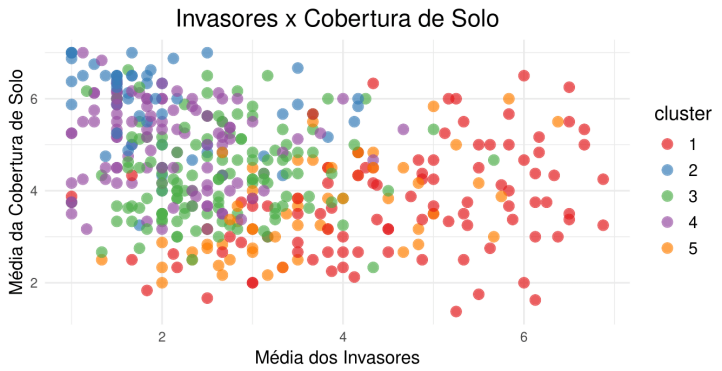

Figure 14: *Scatterplot* envolvendo "CobertSolo" e "Invasoras" coloridas pelas classes propostas. O *scatterplot* mostra uma maior complexidade da classificação proposta, dado que houve uma perturbação maior por outras variáveis.

Conforme se observa pela 14 a nova classificação possui influências diferentes

Então fizemos o mesmo teste para observar o impacto das variáveis sobre a clusterização, utilizamos o mesmo método envolvendo XGBoost e varIMP para avaliação da importância das variáveis preditoras.

xgbTree variable importance	
	Overall
Degrad_media	100.000
DispForr_media	95.585
PotProd_media	79.725
Altura	63.957
Manejo_media	40.545
DispFolhVerd_media	19.591
CondAtual_media	18.879
Invasoras_media	16.280
CobertSolo_media	3.318
Cupins_media	1.382
EstDesenv_media	0.000

xgbTree variable importance	
	Overall
Manejo_media	100.000
CondAtual_media	78.875
DispForr_media	57.197
PotProd_media	28.630
Altura	21.224
Degrad_media	18.305
Cupins_media	16.560
Invasoras_media	4.300
DispFolhVerd_media	3.543
EstDesenv_media	2.574
CobertSolo_media	0.000

xgbTree variable importance	
	Overall
DispForr_media	100.000
Degrad_media	77.785
PotProd_media	44.537
Manejo_media	41.774
CondAtual_media	35.590
Altura	34.167
Invasoras_media	27.656
DispFolhVerd_media	27.541
CobertSolo_media	25.092
EstDesenv_media	9.013
Cupins_media	0.000

xgbTree variable importance	
	Overall
CondAtual_media	100.0000
Degrad_media	87.3366
DispForr_media	50.5254
Invasoras_media	25.7311
Manejo_media	25.1313
Altura	20.1927
DispFolhVerd_media	14.5727
EstDesenv_media	11.0348
CobertSolo_media	2.5728
PotProd_media	0.9617
Cupins_media	0.0000

Figure 15: Valores varIMP resultantes do modelo XGBoost treinado para as classes proposta por nós, indicando uma contribuição mais igualitária entre as covariáveis.

Observe que, diferentemente do método anterior, agora um maior número de variáveis exerce impacto no modelo, como mostrado na Figura 15. Isso é uma vantagem significativa, pois ao incorporar mais variáveis, as classes podem capturar interações mais complexas e fornecer uma visão mais enriquecedora dos dados. Essa abordagem reduz o risco de viés, melhora a capacidade de generalização do modelo e aumenta sua robustez em diferentes cenários.

4 Análise de Concordância

Avaliar a concordância entre avaliadores é essencial para garantir a confiabilidade e a validade das avaliações no estudo em questão. Quando diferentes avaliadores analisam o mesmo fenômeno, espera-se que suas opiniões estejam alinhadas, dado que foram devidamente treinados, refletindo um consenso sobre os critérios de avaliação utilizados.

Inicialmente, optamos por usar testes não paramétricos para avaliar a concordância dos avaliadores em cada experimento. No entanto, a presença de dados faltantes, experimentos com apenas um avaliador e observações registradas incorretamente na base de dados tornaram inviável a aplicação desses testes. Diante disso, optamos por uma abordagem descritiva e visual.

Como os grupos de avaliadores variavam ao longo do processo de amostragem — ou seja, a composição dos grupos mudava para cada ponto de avaliação — decidimos avaliar a concordância de forma temporal. Utilizamos métricas como desvio padrão e amplitude para cada variável ao longo do tempo. A visualização gráfica revelou que o grau de concordância entre os avaliadores era relativamente baixo no início, apresentou uma melhora considerável ao longo do experimento e voltou a piorar no final. Apesar disso, de forma geral há indícios consideráveis de uma boa concordância, visto que, observações muito discrepantes são raras no conjunto de dados.

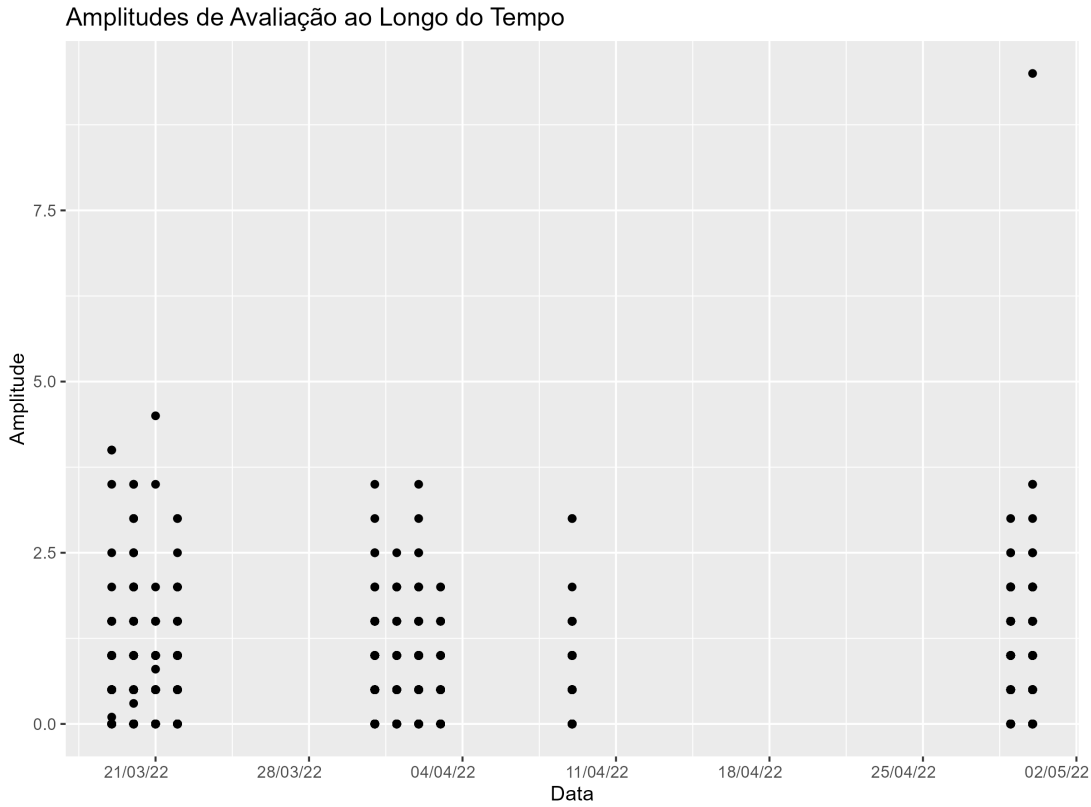

Figure 16: Amplitude geral dos escores por data

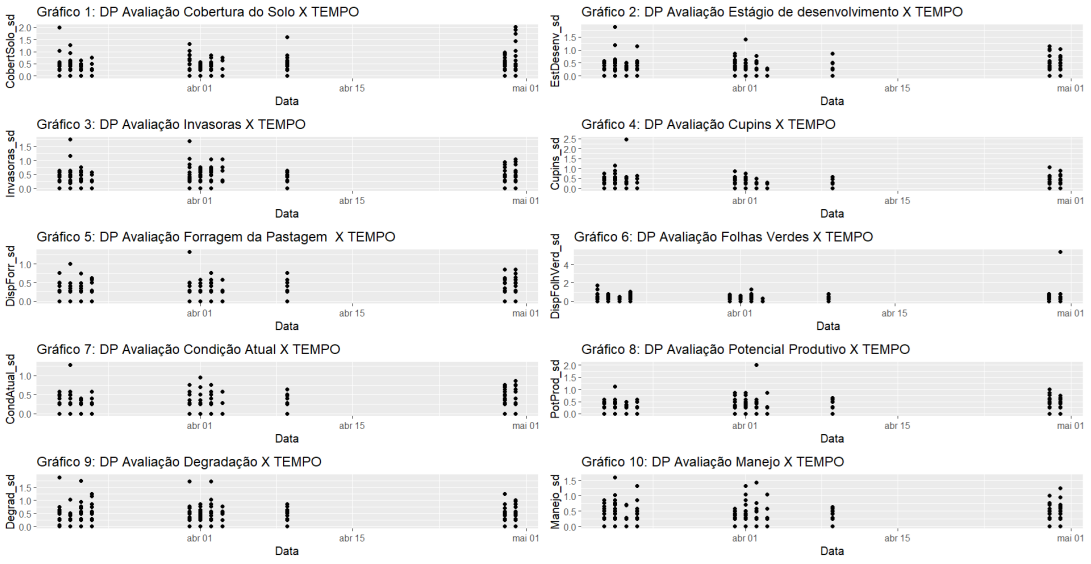

Figure 17: Desvio padrão das avaliações para os diferentes escores ao longo do tempo.

Através da 17 é possível observar a variação do desvio padrão das avaliações dos diferentes escores ao longo do tempo. Observa-se uma diminuição do desvio padrão inicialmente, indicando uma maior concordância entre os avaliadores. Isso pode sugerir que, com o tempo, eles alinharam suas percepções e critérios de avaliação. No entanto, após um certo intervalo sem avaliações, nota-se um aumento nos desvios padrões. Essa elevação pode ser explicada pela pausa nas avaliações ou na mudança dos grupos de avaliadores para cada ponto amostral.

5 *Scripts* em *R* e *Python* que suportam os resultados

Os *scripts* utilizados para gerar o aplicativo web, os gráficos interativos e todas as análises e modelagens relevantes citadas neste documento podem ser encontrados no repositório do GitHub criado pela equipe. O link é: https://github.com/cego669/DatathonEngopeVI.

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Chen, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4).

Kurita, T. (2019). Principal component analysis (pca). *Computer vision: a reference guide*, 1–4.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*.