Atividade 2 MO810/MC959

[*] O link contendo o jupyter notebook e outros conjuntos de dados e também o datacar estão no drive:

https://drive.google.com/drive/folders/1XbJe1t7qfeV_G1yKNZamV8UuZNwLuPZP?usp=sharing

1st Betania E R da Silva
IC-Unicamp

2nd Decio Miranda Filho
IC-Unicamp

3rd Felipe Scalabrin Dosso
IMECC-Unicamp

I. Introdução

Neste trabalho, daremos continuidade à análise do *Adult dataset*, amplamente utilizado em estudos que exploram a desigualdade de renda e buscam identificar vieses em modelos preditivos. O objetivo central é explorar o desempenho de diferentes modelos de aprendizado de máquina e avaliar seu comportamento tanto em termos de desempenho preditivo quanto em relação à justiça, utilizando métricas específicas de *fairness*.

Foram selecionados quatro modelos de classificação: Naive Bayes, K-Nearest Neighbors (KNN), SVM Linear e Random Forest. Cada um desses algoritmos foi escolhido devido à sua relevância e ampla utilização na literatura.

Para avaliar o desempenho preditivo, utilizamos métricas como a acurácia e a *AUC-score*. No entanto, dado o contexto social dos dados e a importância de considerar o impacto de decisões automatizadas em grupos demográficos distintos, também incorporamos métricas de *fairness*, como o *Statistical Parity Difference* (SPD) e o *Disparate Impact* (DI). Essas métricas são essenciais para verificar se os modelos estão reproduzindo desigualdades preexistentes ou se estão apresentando resultados mais justos entre grupos protegidos, como raça e gênero.

Além da avaliação tradicional, realizamos uma otimização dos hiperparâmetros dos modelos, focando em maximizar a *AUC-score*.

O restante do trabalho está organizado da seguinte forma: Seção II discute os trabalhos relacionados, Seção III aborda a metodologia utilizada para a escolha das métricas e modelos utilizados, Seção IV descreve os resultados e por fim, a Seção V discute o resultados encontrados sobre as implicações do uso de tais modelos em contextos sensíveis.

II. Trabalhos Relacionados

Diversos estudos têm explorado diferentes algoritmos de aprendizado de máquina utilizando o *Adult dataset*. A seguir, apresentamos uma revisão de trabalhos relevantes e os modelos aplicados.

Bekena [1] concentrou-se no uso do Random Forest para prever os níveis de renda de indivíduos, obtendo uma acurácia de 85%. O autor destaca que, embora os dados não reflitam os rendimentos atuais, o modelo é útil para identificar os principais fatores que explicam as diferenças entre altos e baixos salários.

Topiwalla [2] implementou diversos modelos, incluindo Decision Tree (com e sem cross-validation), Naive Bayes, Regressão Logística, k-NN, SVM, Random Forest (com e sem cross-validation), XGBoost (com diferentes configurações) e técnicas de stacking como Logistic Stack on XGBoost e SVM Stack on Logistic. Os melhores desempenhos foram observados com Random Forest (sem o atributo *Country*), com 86,89% de acurácia, e XGBoost, com 87,53% de acurácia e AUC de 0,9275.

Lazar [3] investigou o impacto da redução de dimensionalidade via PCA (Análise de Componentes Principais) no desempenho do SVM. O estudo revelou que a acurácia variou pouco, com 84,92% sem redução e 84,42% após a aplicação do PCA (com seis componentes). No entanto, o tempo de execução foi significativamente reduzido, de 932 segundos para 377,7 segundos.

Lemon et al. [4] identificaram as variáveis mais relevantes para prever se a renda de um indivíduo excede $50.000, destacando idade, educação, horas por semana, ocupação e sexo. A Árvore de Decisão obteve o melhor desempenho entre os modelos testados, enquanto o k-NN não conseguiu produzir resultados em tempo hábil, e a Regressão Logística teve dificuldades devido à natureza não linear dos dados.

Por fim, Chakrabarty e Biswas [5] implementaram o Gradient Boosting Classifier (GBC) para prever os níveis de renda, relatando uma acurácia de 88,16% no conjunto de validação.

III. Metodologia

A. Modelagem

A partir das realizações do projeto anterior, em que houve a sanitização, *data-wangling*, codificação *one-hot* para as features categóricas que não entrariam no grupo sensível- (isto é, nas features tidas como sensíve, raça,gênero e país de origem, foram categorizadas de diretamente forma binária). Além disso, com a aplicação do *Min-Max Scaler* e já tendo sido feito o *split* dos dados de treino e de validação - os dados de teste foram preservados dessa análise - iniciar-se-á a fase de modelagem.

1) Métricas de Avaliação: Com a base de dados devidamente limpa e pré-processada, iniciaremos a fase de modelagem.

B. Métricas e Tomada de Decisão

As métricas de avaliação utilizadas foram ***AUC-score (Area Under Curve-score)***, ***Brier-Score*** e apenas como medida de comparação decidiu-se incluir a acurácia. Complementarmente, decidiu-se usar a ***Statistical Parity Diference (SPD)*** e também o ***Disparate Impact (DI)***para observar o impacto dos modelos considerando ***fairness***.

O roc-AUC-score ou AUC é amplamente utilizado para conjuntos de dados desbalanceados por diversos trabalhos científicos [6] [7] [8] [9] [10] muito por conta de efetivamente capturar a área sobre a curva ROC. O ***Brier-Score*** também foi adotado por conta da sua capacidade de quantificar a precisão da probabilidade prevista em relação ao resultado real, ou seja, avalia o quanto as previsões de probabilidade estão distantes dos resultados observados, sobremaneira utilizado em previsões probabilísticas (binárias ou multiclasse) [11] [12] [13]. As métricas de *fairness* adotadas, SPD e *Disparate Impact* (DI), são recomendadas pelo próprio Varshney (2022) [14].

O SPD é definido como:

$$\text{SPD} = P(\hat{y}(X) = \text{fav} \mid Z = \text{desfav}) - P(\hat{y}(X) = \text{fav} \mid Z = \text{priv})$$

Por sua vez, o *Disparte Impact*, de forma análoga ao SPD quantifica a mede a diferença na probabilidade de resultados favoráveis entre grupos desfavorecidos e privilegiados por meio de um *ratio*. Neste, Grupos privilegiados e desprivilegiados são calculados, independentemente da intenção do tomador de decisão e do procedimento de tomada de decisão [14].

1) Procedimentos para Modelagem: Os modelos escolhidos para o processo foram:

- Naive-Bayes (plug-in)
- KNN (plug-in)
- Suport Vector Classifier-Linear (Risk Minimization)
- Random-Forest (Risk Minimization)

Dessa forma, os 2 primeiros são do estilo plug-in, um com a região de separação mais suave e outro menos. E nos 2 últimos o SVC-linear com sepração linear e o Random-Forest, com a capacidade de criar margen de separação mais complexas.

2) Tunning de Hiperparâmetros e Escolha do Threshold de Decisão: A partir da separação realizada na etapa anterior dados de treino e validação por **holdout** deciciu-se tunar os hiperprâmetros pelo framework otimizador **Optuna** o qual foi apresentado por Akiba et al.(2019) [15], utilizando o princípio *define-by-run*, permitindo a criação dinâmica do espaço de busca de parâmetros.

Dessa forma utilizou-se uma malha de hiperparâmetros dos modelos.

Vale notar também que que a função objetivo do Optuna foi otimizada para maximizar a métrica de **AUC-score** e foram computados **50** ***trials*** para cada um dos modelos. Treinados e "testados"sobre o conjunto de validação (Mais detalhes na seção de resultados).

Modelo	Hiperparâmetro	Intervalo ou Valor
RandForestClassif.	n_estimators	(100, 500)
	max_depth	(10, 30)
	min_samples_split	(2, 10)
	min_samples_leaf	(1, 4)
SVC-Linear	C	(0.1, 100)
	penalty	['l2']
	loss	['hinge', 'squared_hinge']
	max_iter	(1000, 3000)
KNeighborsClassif.	n_neighbors	(3, 9)
	weights	['uniform', 'distance']
	p	[1, 2]
GaussianNB	var_smoothing	(1e-9, 1e-7)

Tabela I
MALHA DE HIPERPARÂMETROS

No processo de escolha do *threshold* consideramos a ponderação dos custos e proporção dos falsos positivos e negativos, da seguinte forma,

$$\eta = \frac{c_{01} \cdot p_0}{c_{10} \cdot p_1}$$

em que p_0 e p_1 foram escolhidos com base nas proporções diretas dos dados de treino e os custos foram testadas algumas configurações e decidiu-se penalizar o os erros associados aos Falsos Negativos (FN) 10 vezes mais que o os Falsos positivos (FP)-100:10. Note que se tratando da previsão em que a classe 1 é a população com renda menor que US$50.000 o custo associado ao FN deve ser mais penalizado. Com isso obteve-se o threshold de **0.301582** para calibrarmos nossas probabilidades.

A partir desse processo, foi selecionado o modelo com a melhor configuração de hiperparâmetros e que foram salvos para análise dos resultados.

IV. RESULTADOS

Ao analisar a tabela IV é possível notar que o modelo Random Forest obteve o melhor resultado nas três métricas, com altos valores de AUC (0.9118) e acurácia (0.8340), indicando um ótimo ajuste e alta capacidade de predições sólidas. Além disso, o baixo valor do *Brier Score* sugere que as probabilidades estimadas estão bem calibradas, tornando o modelo confiável para decisões com base nas probabilidades preditas.

Em contraste, o modelo de Naive Bayes apresenta relativamente alta acurácia (0.7973), AUC inferior ao método anterior (0.7516) e *Brier Score* ligeiramente maior (0.1823). Tais métricas sugerem que apesar do modelo ter desempenho razoável em termos de acurácia, sua habilidade de diferenciar entre as classes é limitada e suas probabilidades preditas são menos confiáveis que as de Random Forest.

Assim como o modelo de Naive Bayes, o algoritmo de K-Nearest Neighbors exibe AUC e accuracy similares, refletindo a baixa performance em distinguir entre $> 50K$ e $\leq 50K$. Além do mais, o relativamente alto valor de *Brier*

Score reforça que o KNN não é ideal para essa tarefa de classificação, já que o método tem dificuldade com predição e calibração de probabilidade.

A técnica de Support Vector Classifier (SVC), demonstra acurácia muito inferior aos demais modelos. Este desempenho, junto ausência de informação sobre as suas probabilidades estimadas, demonstra que o modelo SVC não é apropriado para esse conjunto de dados. Sua abordagem linear provalvelmente não consegue capturar as complexidades da fronteiras de decisão necessárias para uma classificação eficaz.

Modelo	**AUC**	**Acurácia**	**Brier Score***
Random Forest (rf)	0.9118	0.8340	0.1002
Naive Bayes (nb)	0.7973	0.7516	0.1823
K-Nearest Neighbors (knn)	0.6159	0.7507	0.2020
Support Vector Classifier (SVC)	0.5000*	0.2484	N/A

Tabela II
COMPARAÇÃO DAS MÉTRICAS DE AUC, ACURÁCIA, E BRIER SCORE ENTRE OS MODELOS.

*Para o modelo SVC, como não há valor de AUC disponível, utilizamos a acurácia balanceada.

	Base (Sem Modelo)		Random Forest		Naive Bayes	
Métrica	SPD	DI	SPD	DI	SPD	DI
Sexo	0.19	0.36	0.09	0.30	0.00	0.00
Raça	0.10	0.60	0.06	0.44	0.00	0.00
País Nativo	0.05	0.76	0.05	0.47	0.00	0.00
	K-Nearest Neighbors		Support Vector Classifier			
Métrica	SPD	DI	SPD	DI		
Sexo	0.05	0.32	0.00	1.00		
Raça	0.03	0.47	0.00	1.00		
País Nativo	0.028	0.54	0.00	1.00		

Tabela III
COMPARAÇÃO DAS MÉTRICAS DE *fairness* (SPD E DI) ENTRE O CASO BASE E OS MODELOS TREINADOS.

Para a análise de *fairness*, em que calculados o SPD e DI, nota-se que para a base inicial (sem predições do modelo) há um viés demonstrado pelo SPD, conforme a tabela III, nota-se que para os três atributos protegidos (Sexo feminino, indivíduos de raça negra e país nativo fora dos EUA).

Após termos obtido as previsões do modelo percebemos variações do SPD. No Random Forest percebemos que que a disparidade para raça e sexo melhorou, apesar de ainda continuar mostrando viés. Acredita-se que tanto as métricas de desempenho terem sido mais elevadas que em relação aos outros modelos tenha sido um fator para isso. Em contrapatida, o KNN mostrou uma sensível redução do SPD/DI, mas ao mesmo tempo demonstrou métricas de performance não tão elevadas. Já os algoritmos de SVC-Linear e Naive-Bayes tiveram o spd "zerados", mas por conta do poder de predição do modelo ter sido insatisfatório. Note aqui que a acurácia do SVC com fronteira linear ter sido de 50% demonstra isso.

V. DISCUSSÃO

Portanto, dada a expoisção feita sobre os resultados nota-se que nosso melhor modelo- tanto em termos de métricas de performance quanto em métricas de *fairness* foi o Random Forest apresentando uma alta capacidade - tanto de acurácia (83%) e AUC (91%) quanto da métrica de SPD.

O modelo Naive-Bayes mostrou-se com resultados intermediários enquanto o SVC-Linear apresentou-se com o ior desempenho. Nesse caso, acredita-se que a margem linear e complexidade do algortimo tenham sido insuficientes para separar os dados, e neste caso a otimização de hiperparâmetros tenha sido incapaz de melhorar o modelo. Além disso, o KNN mostrou-se com um desempenho também intermediário, mas levanta-se a hipótese de que testá-lo com configurações mais robustas poderia levá-lo a melhores resultados.

Dessa forma, pontos a serem reconsiderados seriam tornar os modelos com fronteiras mais robustas (no caso do SVC) e testar mais configurações de hiperparâmetros. Acredita-se que testar em outras variedades de modelos traria um ganho na diversidade da análise.

A análise de fairness, feita de forma preliminar, há de ser aprofundada e melhor avaliada em etapas posteriores.

REFERÊNCIAS

[1] S. M. Bekena, "Using decision tree classifier to predict income levels," *Munich Personal RePEc Archive*, July 30 2017, mPRA Paper No. 80816.

[2] M. Topiwalla, "Machine learning on uci adult data set using various classifier algorithms and scaling up the accuracy using extreme gradient boosting," 2017, unpublished manuscript.

[3] A. Lazar, "Income prediction via support vector machine," in *International Conference on Machine Learning and Applications (ICMLA)*, Louisville, KY, USA, December 16–18 2004.

[4] C. Lemon, C. Zelazo, and K. Mulakaluri, "Predicting if income exceeds $50,000 per year based on 1994 us census data with simple classification techniques," https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/048.pdf, 2015.

[5] N. Chakrabarty and S. Biswas, "A statistical approach to adult census income level prediction," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018, pp. 207–212.

[6] J. seok Lee, "Auc4.5: Auc-based c4.5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106 034–106 042, 2019.

[7] Asniar, N. Maulidevi, and K. Surendro, "Smote-lof for noise identification in imbalanced data classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 3413–3423, 2021.

[8] B. Zhou, Y. Ying, and S. Skiena, "Online auc optimization for sparse high-dimensional datasets," *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 881–890, 2020.

[9] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical roc-auc estimation in binary classification tasks," *Journal of Statistical Computation and Simulation*, vol. 94, pp. 183 – 203, 2023.

[10] Y.-C. Wang and C.-H. Cheng, "A multiple combined method for rebalancing medical data with class imbalances," *Computers in biology and medicine*, vol. 134, p. 104527, 2021.

[11] H. Kvamme and Ørnulf Borgan, "The brier score under administrative censoring: Problems and solutions," *J. Mach. Learn. Res.*, vol. 24, pp. 2:1–2:26, 2019.

[12] A. H. Murphy, "A new decomposition of the brier score: formulation and interpretation," *Monthly Weather Review*, vol. 114, pp. 2671–2673, 1986.

[13] D. Stephenson, C. A. S. Coelho, and I. Jolliffe, "Two extra components in the brier score decomposition," *Weather and Forecasting*, vol. 23, pp. 752–757, 2008.

[14] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019.