Análise Comparativa de Modelos de Classificação Aplicados a uma Base de Dados real

Igor Ryan Bacelar Frota[1], **Samuel Silva dos Santos**[1]

[1]Instituto Federal do Maranhão (IFMA) – *Campus* Coelho Neto
Coelho Neto – MA – Brasil

{igorb,santos.samuel1}@acad.ifma.edu.br

Abstract. *This work presents a comparative analysis of supervised classification models applied to the Adult Income dataset, aiming to predict whether an individual's annual income exceeds fifty thousand dollars based on demographic attributes. The models used were: Multilayer Perceptron (MLP), K-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest (RF). The study involves data preprocessing, application of algorithms, and performance evaluation using metrics such as accuracy, precision, recall, and F1-Score.*

Resumo. *Este trabalho apresenta uma análise comparativa de modelos de classificação supervisionada aplicados à base de dados 'Censo de Renda Adulta' (Adult Census Income), com o objetivo de prever se a renda de um indivíduo ultrapassa 50 mil dólares anuais com base em atributos demográficos. Os modelos utilizados foram: Multilayer Perceptron (MLP), K-Nearest Neighbors (kNN), Support Vector Machine (SVM) e Random Forest (RF). O estudo envolve pré-processamento dos dados, aplicação dos algoritmos e avaliação de desempenho por métricas como acurácia, precisão, recall e F1-Score.*

1. Introdução

A classificação supervisionada é uma técnica amplamente utilizada em problemas reais de previsão e tomada de decisão[Ferreira et al. 2021]. O presente trabalho tem como objetivo avaliar o desempenho de quatro modelos de classificação – k-Vizinhos Mais Próximos (kNN), Máquina de Vetores de Suporte (SVM), Florestas Aleatórias (RF) e Rede Neural Multicamadas (MLP) – aplicados à base 'Censo de Renda Adulta' (*Adult Census Income*), um conjunto de dados com atributos demográficos e profissionais, no intuito de prever a faixa de renda dos indivíduos. Utiliza-se como referência a base de dados disponível no repositório da *UCI Machine Learning Repository* [Dua and Graff 2017].

A aprendizagem de máquina, como definido por [Lenz et al. 2020],"é um subcampo da inteligência artificial que permite aos sistemas aprenderem padrões a partir de dados sem programação explícita". Para que os algoritmos de aprendizado consigam performar satisfatoriamente sobre esses dados faz-se necessária a correta preparação da base, o que ocorre na etapa de análise exploratória dos dados (EDA) sendo esta uma etapa crítica, pois, como destacam [Ferreira et al. 2021], "a qualidade das conclusões depende diretamente da preparação inicial dos dados".

2. Descrição da Base de Dados

A base 'Censo de Renda Adulta', oriunda do repositório UCI, contém 48842 instâncias e 14 atributos preditivos. A variável-alvo é categórica, indicando se o indivíduo possui renda superior a 50 mil dólares anuais ("$>$50K") ou não ("$\leq$50K"). Os atributos incluem idade, nível de escolaridade, ocupação, horas trabalhadas por semana, entre outros (Tabela 1). O conjunto de dados apresenta valores ausentes representados por "?" e variáveis categóricas que requerem codificação e um pré-processamento adequado para o uso dos algoritmos de aprendizado de máquina. Por não apresentar alta capacidade preditiva, esta base se mostra adequada para fins experimentais, permitindo avaliar o desempenho e a robustez de diferentes algoritmos de aprendizado de máquina

Tabela 1. Principais Características da Base de Dados

Característica	Descrição
Fonte	*UCI Machine Learning Repository (ID 2)*
Tamanho	48.842 instâncias
Atributos	14 (6 numéricos, 8 categóricos)
Variável alvo	Renda anual (binária: $> 50K$ ou $\leq 50K$)
Desbalanceamento	23% na classe $> 50K$
Atributos numéricos	idade, anos de educação, ganho/perda de capital, horas/semana
Atributos categóricos	ocupação, educação, estado civil, raça, sexo, país de origem
Dados faltantes	Presentes (5.7% em "ocupação"e "tipo de emprego")
Aplicação típica	Classificação binária, estudos de desigualdade social
Desafios	Viés em atributos sensíveis (gênero, raça), dados desbalanceados

Fonte: Adaptado de [Lichman 2013].

3. Metodologia

3.1. Etapas Adotadas

As principais etapas adotadas na abordagem foram:

- **Pré-processamento**: Foram utilizadas duas codificações diferentes como sugerido por [Lenz et al. 2020], aplicou-se o *One-Hot Encoding* aos atributos categóricos uma vez que, conforme apontado por [learn Developers 2023], este último configura-se como mais adequado para algoritmos baseados em distancia euclidiana como é o caso do kNN e SVM.
- **Seleção de Modelos**: Foram testados quatro algoritmos, baseados nas recomendações de [Mariano et al. 2020] e [Ferreira et al. 2021]; Sendo os mesmos o kNN, SVM, RF e MLP.
- **Treinamento**: Cada algoritmo após a etapa de pré-processamento foi submetido a 10 testes consecutivos através da técnica de validação cruzada conforme o padrão ouro exposto por [Lenz et al. 2020].
- **Avaliação**: Utilizou-se validação cruzada,para amenizar o *overfiting*, método defendido por [Lenz et al. 2020] para evitar otimismo indevido, assim como outras métricas de avaliação como Acurácia, Precisão, *Recall*, *F1-Score* e matriz de confusão. As principais métricas de avaliação a ser consideradas nos testes foi a Acurácia, *Recall* e *F1-Score* seguindo as orientações sugeridas por [Murphy 2012].

3.2. Pré-processamento

- Tratamento de valores faltantes: Optou-se pela remoção dos valores representados por interrogação (?), uma vez que sua quantidade era pouco expressiva e seu tratamento não traria impacto relevante na qualidade dos dados.
- Tratamento de atributos categóricos: Para codificação dos atributos categóricos foi utilizado o *One-Hot Encoding* tendo em vista que tal codificação melhor se configura para algoritmos baseados em distância como é o caso do SVM e kNN além de padronizar a codificação destes atributos para todos os algoritmos.
- Normalização *Min-Max* e *Stander-Scaler* para atributos numéricos.
- Divisão dos dados: separação em 70% para treinamento e 30% para teste.

3.3. Modelos

Visando uma análise do desempenho bruto dos algoritmos, optou-se por não aplicar ajustes de hiperparâmetros ou técnicas de otimização, exceto pela validação cruzada, utilizando-os com as configurações padrão conforme descritas em suas documentações. A MLP (rede neural) foi submetida a ajustes e testes preliminares até atingir o melhor desempenho possível antes da etapa de avaliação final.

- **MLP**: 3 camadas ocultas (200,100,50 neurônios respectivamente), ReLU
- **kNN**: k=5 (otimizado por validação cruzada)
- **SVM**: Kernel RBF, C=1.0
- **RF**: 100 árvores, max_depth=10

O desenvolvimento foi realizado no *Google Colab* com as bibliotecas `scikit-learn` [Pedregosa et al. 2011], `pandas`, `numpy` e `keras` . A base teórica sobre redes neurais foi fundamentada em [Bishop 2006].

4. Resultados

Os modelos foram avaliados pelas métricas: acurácia, precisão, *recall* e *F1-Score*. Os quais serão detalhados a seguir.

A Tabela 2 a presenta o desempenho dos modelos em relação à acurácia média resultante da validação cruzada com uma padronização de 10 (dez) *folds*. É possivel identificar que a maioria dos modelos performaram relativamente bem obtendo acurácia acima de 0.85 à exceção do kNN cuja acurácia ficou um pouco abaixo.

Tabela 2. Desempenho dos Modelos em relação à Acurácia Média

Modelo	Acurácia Média
MLP	0.8507
kNN	0.8295
SVM (RBF)	**0.8557**
RF	0.8542

Ao se avaliar as métricas de Precisão, *Recall* e *F1-Score* para predição de renda inferior ou igual a 50 mil dólares anuais, os modelos performaram conforme apresentado na Tabela 3.

Tabela 3. Desempenho dos Modelos na Predição de Renda inferior ou igual a 50 mil dólares anuais

Modelo	Precisão	*Recall*	*F1-Score*
MLP	0.89	0.91	0.90
kNN	0.87	0.90	0.89
SVM (RBF)	0.87	0.95	0.91
RF	0.88	0.93	0.90

Tabela 4. Desempenho dos Modelos na Predição de Renda Superior a 50 mil dólares anuais

Modelo	Precisão	*Recall*	*F1-Score*
MLP	0.71	0.67	0.69
kNN	0.68	0.60	0.64
SVM (RBF)	0.78	0.57	0.65
RF	0.74	0.61	0.67

A Tabela 4 apresenta os valores de Precisão, *Recall* e *F1-Score* para predição de renda superior a 50 mil dólares anuais.

Ao se analisar as Tabelas 3 e 4 observa-se que todos os modelos desempenham melhor para a predição de renda anual menor ou igual a 50 mil dólares anuais algo que se deve ao fato da base objeto de estudo ser desbalanceada contendo um percentual de 76,07% da classe $\leq 50K$.

A Figura 1 apresenta o desempenho do algoritmo MLP que conseguiu bons resultados, porém com maior custo computacional quando comparado com os demais.

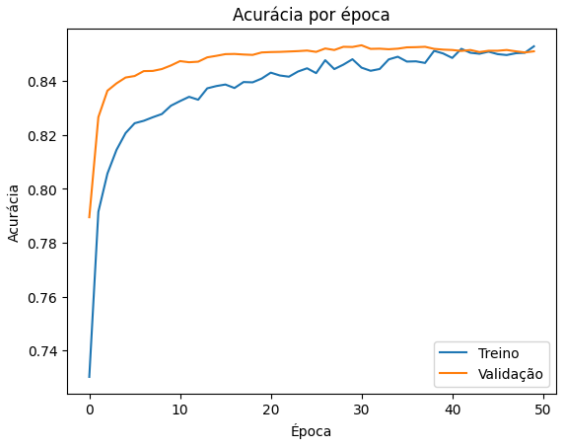

Figura 1. Acurácia por época com MLP

5. Discussão

Como destacado por [Ferreira et al. 2021] e também por [Lenz et al. 2020] a preparação inicial dos dados configura-se como crucial para a qualidade das conclusões resultantes da aplicação dos modelos de aprendizagem de máquina.

O pré-processamento teve papel fundamental na performance dos modelos. Modelos baseados em distância, como kNN principalmente, pode ser ser especialmente sensível à normalização e/ou codificação de variáveis categóricas uma vez que este modelo, assim como o SVM, são baseados em distância e a codificação inadequada de atributos categóricos pode enviesar o modelo.

A base objeto de estudo possui um moderado grau de desbalanceamento (proporção aproximada de $\sim 3.2:1$) o que pode impactar significativamente na avaliação dos modelos. [Lenz et al. 2020] recomenda que em bases com essas características a métrica *F1-Score* seja priorizada por ser mais robusta uma vez que procura balancear a Precisão e o *Recall*, tendo em vista que a Precisão ignora os falsos negativos e o *Recall* ingnora os falsos positivos.

6. Conclusão

O estudo comparativo evidenciou a eficácia de todos os modelos na tarefa proposta, de forma que todos os modelos performaram relativamente bem. Além disso, reforçou-se a importância do pré-processamento na obtenção de resultados mais consistentes e precisos.

Como o objetivo foi avaliar o desempenho dos modelos com a configuração padrão dos algoritmos conforme sua documentação [Pedregosa et al. 2011] sem nenhum ajuste adicional como, por exemplo, ajustes de hiperparâmetros, dentre outros, esses foram os resultados obtidos. A fim de se melhorar tais resultados, em trabalhos futuros pode-se incluir técnicas avançadas de redução de dimensionalidade (ex.: PCA) bem como técnicas de balanceamento de classes (ex.: SMOTE) para otimização conforme sugerido por [Ferreira et al. 2021], tendo em vista que a codificação *One-Hot* pode aumentar significativamente essa dimensionalidade.

Alumas técnicas de mineração de regras de associação, abordadas em [Mariano et al. 2020] podem ser aplicadas a fim de enriquecer a análise, como por exemplo o algoritmo apriori que é capaz de identificar itens frequentes e regras no tipo "Se X, então Y"(ex.: se `educação=Universitário, ocupação=Gerente` então `renda>50K`).

Como sugerido por [Lenz et al. 2020], trabalhos futuros poderiam explorar técnicas de Aprendizado por Transferência (*transfer learning*) para otimizar o treinamento do MLP, tendo em vista que a base Censo de Renda Adulta possui aproximadamente 48,8 mil atributos o que pode incorrer em um *overfitting* a implementação de um MLP tradicional. O Aprendizado por Transferência permitiria começar com um modelo já "educado".

Referências

[Bishop 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

[Dua and Graff 2017] Dua, D. and Graff, C. (2017). Uci machine learning repository: Adult data set. Acesso em: jul. 2025.

[Ferreira et al. 2021] Ferreira, R. G. C., Miranda, L. B. A., Pinto, R. A., Pessutto, L. R. C., Pereira, M. A., Andrade, A. L. C. d., and Marques, L. T. (2021). *Preparação e análise exploratória de dados*. SAGAH, Porto Alegre. [recurso eletrônico].

[learn Developers 2023] learn Developers, S. (2023). *Preprocessing Data: Encoding Categorical Features*. Seção: ”Encoding categorical features”.

[Lenz et al. 2020] Lenz, M. L., Neumann, F. B., Santarelli, R., and Salvador, D. (2020). *Fundamentos de aprendizagem de máquina*. SAGAH, Porto Alegre. [recurso eletrônico].

[Lichman 2013] Lichman, M. (2013). UCI machine learning repository. Dataset: Adult Census Income (ID 2).

[Mariano et al. 2020] Mariano, D. C. B., Marques, L. T., Silva, M. S., Júnior, J. F. M. A., Santos, M. d. S. d., and Santos, T. m. d. O. (2020). *Data mining*. SAGAH, Porto Alegre. [recurso eletrônico].

[Murphy 2012] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 1st edition.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. Acesso em: jul. 2025.