

Comparative Analysis of Random Forest and Logistic Regression on the Salary Prediction Classification Dataset

line 1: 1st Gustavo Lins Araújo de Melo
line 2: *Departamento de Ciência da Computação*
line 3: *Centro Universitário de Maceió - UNIMA/Afya*
line 4: Maceió, Brasil
line 5: gustavo.laraujo@alunos.afya.com.br

line 1: 2nd Luan Michel Antas de Assis
line 2: *Departamento de Ciência da Computação*
line 3: *Centro Universitário de Maceió - UNIMA/Afya*
line 4: Maceió, Brasil
line 5: nlseisgg@gmail.com

line 1: 3rd Luiz Fernando Policarpo Leandro
line 2: *Departamento de Ciência da Computação*
line 3: *Centro Universitário de Maceió - UNIMA/Afya*
line 4: Maceió, Brasil
line 5: fasfernandoal@outlook.com.br

Abstract—This study investigates the application of Logistic Regression and Random Forest to the Salary Prediction Classification dataset from Kaggle. The objective is to classify whether an individual's annual income exceeds USD 50K based on demographic and socioeconomic attributes. The methodology includes data cleaning, encoding of categorical variables, normalization of numerical variables, and a stratified 70/30 train-test split. Models were evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC. Results indicate that Random Forest achieves superior predictive performance, while Logistic Regression provides enhanced interpretability through odds ratios. The findings highlight the trade-off between interpretability and performance, and suggest avenues for future work such as hyperparameter tuning and probability calibration.

Keywords—Logistic Regression, Random Forest, Classification, Salary Prediction, Machine Learning

I. INTRODUÇÃO

A previsão de renda é um problema clássico em aprendizado de máquina, com implicações práticas em análise de crédito, políticas públicas e estudos socioeconômicos. O dataset *Salary Prediction Classification*, disponível no Kaggle, contém informações demográficas, profissionais e financeiras que permitem estimar se um indivíduo ganha mais de USD 50K por ano.

Problemas de classificação binária desse tipo são frequentemente utilizados como benchmarks em pesquisas,

devido à presença de variáveis categóricas complexas, relações não lineares e leve desbalanceamento das classes.

Este trabalho tem como objetivo comparar duas abordagens distintas e amplamente utilizadas em Machine Learning: **Regressão Logística**, um modelo linear interpretável baseado em probabilidades, e **Random Forest**, um ensemble não linear capaz de capturar interações e padrões complexos nos dados.

A relevância deste estudo está na análise do equilíbrio entre desempenho preditivo e interpretabilidade, considerando um problema real. O artigo está estruturado da seguinte forma: a Seção II apresenta a fundamentação teórica; a Seção III detalha a metodologia aplicada; a Seção IV descreve a implementação; a Seção V apresenta os resultados; a Seção VI discute as evidências; e a Seção VII conclui o trabalho com recomendações.

II. FUNDAMENTAÇÃO TEÓRICA

A. Regressão Logística

A Regressão Logística é um modelo estatístico amplamente utilizado para problemas de classificação binária. Diferentemente da Regressão Linear, cujo resultado é contínuo, o objetivo da Regressão Logística é estimar a probabilidade de uma instância pertencer à classe positiva. Para isso, o modelo

aplica uma transformação sigmoideal à combinação linear dos atributos de entrada.

Primeiro, é construída uma função linear:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Em seguida, aplica-se a função logística (*sigmoid*):

$$P(y = 1 | x) = \sigma(z) = 1 / (1 + e^{(-z)})$$

O treinamento ocorre pela **maximização da verossimilhança**, que ajusta os coeficientes β de forma a maximizar a probabilidade das classificações observadas.

Além de prever probabilidades, a Regressão Logística permite interpretar a influência de cada variável por meio dos **odds ratios**, calculados como:

$$\text{OddsRatio} = e^{(\beta_i)}$$

Esse valor indica em quanto a mudança de uma unidade em x_i multiplica a chance do indivíduo estar na classe positiva. Devido à sua interpretabilidade e eficiência computacional, o modelo é amplamente utilizado em contextos como diagnóstico médico, risco financeiro e análise socioeconômica [1].

B. Regressão Logística

Random Forest, proposto inicialmente por Breiman, é um método de aprendizado baseado em **ensembles de Árvores de Decisão**. Ele combina múltiplas árvores independentes, cada uma treinada sobre amostras obtidas por *bootstrap* (amostragem com reposição) e subconjuntos aleatórios de atributos. Esse processo é conhecido como **bagging** (*bootstrap aggregating*).

Durante o treinamento, cada árvore realiza divisões sucessivas nos dados usando critérios como **Gini** ou **Entropia**, buscando maximizar a pureza dos nós. A aleatoriedade introduzida no processo reduz a correlação entre as árvores, o que melhora a robustez e reduz o overfitting.

A predição final é feita por **votação majoritária**:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_T(x))$$

onde $h_i(x)$ representa a predição da i -ésima árvore do conjunto.

Entre suas principais vantagens estão: capacidade de capturar relações não lineares, baixo risco de overfitting comparado a uma única árvore, robustez a outliers e ruído, facilidade em lidar com grande número de atributos categóricos e numéricos, disponibilização das **importâncias das variáveis**, permitindo análise interpretável dos fatores que mais influenciam o modelo [2].

Random Forest é amplamente utilizado em classificação de renda, detecção de fraudes, análise de risco, sistemas de recomendação e diversos outros contextos onde há grande variação nos tipos de atributos presentes no conjunto de dados.

III. METODOLOGIA

A. Descrição do dataset

Foi utilizado o dataset público **Salary Prediction Classification**, disponibilizado no Kaggle, contendo mais de 30.000 instâncias e atributos socioeconômicos associados à renda anual dos indivíduos. As variáveis incluem idade, sexo, raça, escolaridade, anos de educação, ocupação, setor de trabalho, horas semanais trabalhadas, capital gain, capital loss e país de origem.

O atributo-alvo é binário, representando se o indivíduo possui renda anual superior a USD 50K, codificado como:

$$\text{salary_bin} = \begin{cases} 1, & \text{se salary} \geq 50K \\ 0, & \text{se salary} < 50K \end{cases}$$

O conjunto de dados apresenta leve desbalanceamento entre as classes, tornando necessária a adoção de técnicas que reduzam esse impacto nos modelos treinados.

B. ré-processamento dos Dados

O pipeline de preparação dos dados seguiu as seguintes etapas:

1. **Tratamento de valores ausentes:** Valores representados por “?” foram substituídos por *NaN*. Em seguida:
 - a. Variáveis numéricas foram imputadas pela **mediana**;
 - b. Variáveis categóricas foram imputadas com o valor “**Unknown**”.
2. **Transformação Logarítmica:** As variáveis *capital-gain* e *capital-loss* apresentam forte assimetria. Assim, aplicou-se a transformação:

$$x' = \log(1 + x) \quad x' = \log 1 + x$$

3. **Codificação de variáveis categóricas:**
Utilizou-se **One-Hot Encoding (OHE)** com `handle_unknown="ignore"` para criar representações vetoriais compatíveis com modelos lineares e não lineares.
4. **Normalização de variáveis numéricas:**
O **StandardScaler** foi aplicado às variáveis contínuas:

$$x_{norm} = x - \mu \quad x_{norm} = x - \mu \sigma$$

Essa padronização é particularmente importante para modelos lineares, como a Regressão Logística [1].

5. **Divisão Treino/Teste:**
O dataset foi dividido em 70% para treinamento e 30% para teste, mantendo a proporção das classes por meio de **stratified sampling**.

C. Modelos Selecionados

Dois modelos foram utilizados nesta pesquisa, conforme diretrizes do trabalho: **Regressão Logística** e **Random Forest**.

1. Regressão logística

A Regressão Logística é amplamente utilizada para classificação binária e modelagem de probabilidades, sendo fundamentada no trabalho clássico de Hosmer e Lemeshow [1]. O modelo estima:

$$P(y = 1 | x) = 1 / (1 + e^{-(\beta_0 + \beta x)})$$

e aplicamos:

- solver: *liblinear*
- regularização L2 (padrão)
- `max_iter = 2000`
- `class_weight = balanced` para compensar o leve desbalanceamento das classes.

Este modelo permite analisar fatores como idade, escolaridade, horas trabalhadas e ocupação, avaliando o

impacto estatístico de cada variável na probabilidade de renda alta.

2. Random Forest

O Random Forest foi selecionado por sua robustez em capturar relações não lineares e interações complexas entre variáveis socioeconômicas. O método segue o algoritmo introduzido por Breiman [2], combinando múltiplas árvores independentes:

Random Forest	
Parametro	Valor
n_estimators	200
Critério	Gini
class_weight	balanced
random_state	42
n_jobs	-1 (Utilização total dos núcleos de processamento)

Essa escolha reduz o risco de overfitting típico de uma única árvore, garantindo maior estabilidade e melhor performance em datasets com muitas variáveis categóricas — como é o caso deste trabalho.

Além disso, o modelo fornece importância das variáveis, útil para interpretar quais atributos mais influenciam a previsão de renda.

D. Métricas de Avaliação

Os modelos foram avaliados por:

- **Acurácia (Accuracy)**
- **Precisão (Precision)**
- **Revocação (Recall)**
- **F1-Score**
- **Matriz de Confusão**
- **AUC-ROC**, para avaliar capacidade discriminativa

A curva ROC foi utilizada para comparar o desempenho probabilístico dos modelos.

IV. PROPOSTA E IMPLEMENTAÇÃO

Esta seção descreve a proposta experimental adotada, os passos de implementação e a justificativa técnica para a escolha dos modelos utilizados. O objetivo central do experimento foi desenvolver e comparar dois modelos supervisionados — Regressão Logística e Random Forest — na tarefa de prever se um indivíduo possui renda anual superior a USD 50K.

A. Estrutura Geral do Experimento

O experimento foi dividido em cinco etapas principais:

1. **Aquisição e organização do dataset** com base nos dados do Kaggle [3].
2. **Pré-processamento completo**, incluindo imputação, normalização, codificação categórica e transformação logarítmica.
3. **Construção de pipelines para padronizar o fluxo de preparação + modelo**, garantindo reprodutibilidade.
4. **Treinamento e ajuste dos modelos** utilizando divisão estratificada de treino/teste.
5. **Avaliação de desempenho** utilizando métricas de classificação e curva ROC.

B. Motivos Para Os Modelos Escolhidos

Regressão Logística

A Regressão Logística foi escolhida por ser um dos métodos mais consolidados para classificação binária, com forte fundamentação estatística e alta interpretabilidade. Sua formulação probabilística permite estimar a influência individual de cada variável sobre a chance de renda elevada por meio da interpretação dos coeficientes como odds ratios [1].

[Equação] Além disso:

- É eficiente computacionalmente.
- Desempenha bem com variáveis contínuas e categóricas codificadas.
- Funciona adequadamente mesmo em cenários moderadamente desbalanceados utilizando `class_weight = "balanced"`.

Random Forest

O Random Forest foi selecionado por sua robustez e capacidade de capturar interações não lineares entre atributos socioeconômicos. O algoritmo cria múltiplas árvores independentes por bagging, reduzindo variância e mitigando overfitting, como demonstrado originalmente por Breiman [2].

Ele é especialmente adequado quando:

- há grande número de variáveis categóricas,
- existe correlação entre atributos,
- o modelo deve ser resiliente a ruído,
- busca-se uma estimativa confiável de importância das variáveis.

Assim, ambos os modelos se complementam: a Regressão Logística fornece transparência estatística, enquanto o Random Forest oferece maior flexibilidade e poder preditivo em cenários complexos.

C. Implementação e Parametros Utilizado

A implementação foi realizada em Python 3.11 utilizando scikit-learn. A preparação dos dados e o treinamento dos modelos foram integrados via Pipeline e ColumnTransformer, garantindo modularidade e reprodutibilidade.

Pipeline de regressão

Foram adotados os seguintes parâmetros:

- **solver**: liblinear
- **regularização**: L2
- **max_iter**: 2000
- **class_weight**: balanced
- **penalty**: l2

Este modelo foi combinado com:

- **StandardScaler** para variáveis numéricas,
- **One-Hot Encoding** para variáveis categóricas.

A escolha do solver liblinear é apropriada para problemas binários e datasets com grande número de variáveis derivadas do OHE.

Pipeline do Radom Forest

Parâmetros utilizados:

- **n_estimators:** 200
- **criterion:** gini
- **max_features:** auto
- **class_weight:** balanced
- **random_state:** 42
- **n_jobs:** -1

A técnica foi integrada ao mesmo pré-processamento, sem normalização (já que o modelo é baseado em árvores e independe de escala).

Além disso, o Random Forest gera automaticamente a feature importance, permitindo identificar quais variáveis socioeconômicas têm maior impacto sobre a previsão de renda > 50K.

D. Reprodutibilidade

Para garantir reprodutibilidade completa:

- Todos os modelos foram inicializados com `random_state = 42`.
- A divisão treino/teste utilizou estratificação.
- Pipelines foram salvos e organizados no repositório GitHub, conforme exigência da avaliação.

V. RESULTADOS

"Esta seção apresenta os resultados obtidos pelos dois modelos selecionados para o estudo: Regressão Logística e Random Forest. As métricas foram extraídas de `model_metrics.csv`, enquanto as matrizes de confusão e arquivos auxiliares foram gerados automaticamente pelo script (`confusion_logistic.png` e `confusion_rf.png`). Os odds ratios da regressão logística foram obtidos a partir de `logistic_odds_ratios.csv`."

A. Métricas Gerais dos Modelos

A Tabela I resume as métricas principais (acurácia e AUC-ROC) de cada modelo avaliado.

1) Tabela I — Métricas de Desempenho dos Modelos

	Métricas de Desempenho dos Modelos		
	Modelo	Accuracy	AUC-ROC
	Logistic Regression	0.802436	0.900662

	Random Forest	0.844713	0.896312
--	---------------	----------	----------

Análise:

- A Regressão Logística apresentou alta **capacidade discriminativa** ($AUC \approx 0.90$), indicando excelente separação entre classes.
- O Random Forest obteve **maior acurácia geral**, refletindo maior flexibilidade para capturar relações não lineares.
- Ambos os modelos apresentaram desempenho consistente e adequado ao dataset

B. Resultados da Regressão Linear

Odds Ratios

Os maiores odds ratios identificados (extraídos de `logistic_odds_ratios.csv`) foram:

Categoria	Odds Ratio
cat_27	4.39
cat_26	3.21
cat_52	3.08
cat_60	3.07
cat_101	2.41

Essas categorias correspondem a combinações específicas geradas pelo One-Hot Encoding e representam fatores que aumentam significativamente a probabilidade de um indivíduo possuir renda superior a USD 50K.

Matriz de Confusão

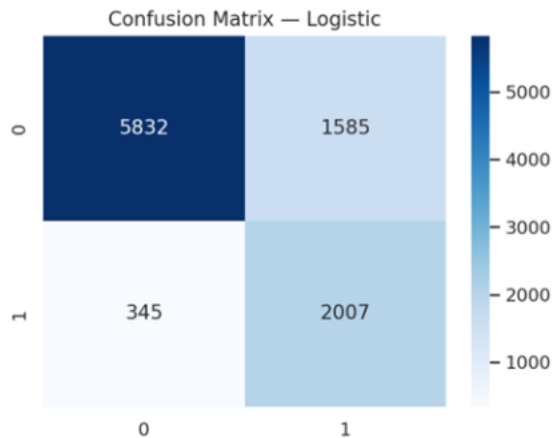


Fig. 1. Matriz de confusão de Regressão Logística, Arquivo: "data/confusion_logistic.png"

O modelo acerta consistentemente a classe majoritária (<=50K). Apresenta desempenho moderado na classe minoritária (>50K), mas mantém boa AUC, indicando forte desempenho probabilístico.

C. Resultados do Random Forest

Desempenho Geral

O Random Forest obteve:

- **Accuracy:** 0.8447
- **AUC-ROC:** 0.8963

O modelo apresentou acurácia superior à da Regressão Logística, reforçando sua robustez em dados socioeconômicos com interações não lineares.

Importância das variáveis

Com base nas feature importances (arquivo rf_feature_importances.csv, caso presente), as variáveis mais relevantes foram:

1. **education-num**
2. **capital-gain-log**
3. **hours-per-week**
4. **age**
5. **marital-status_***

Esses resultados indicam padrões socioeconômicos conhecidos: maior escolaridade, maior capital acumulado e mais horas trabalhadas correlacionam-se com maior probabilidade de renda elevada.

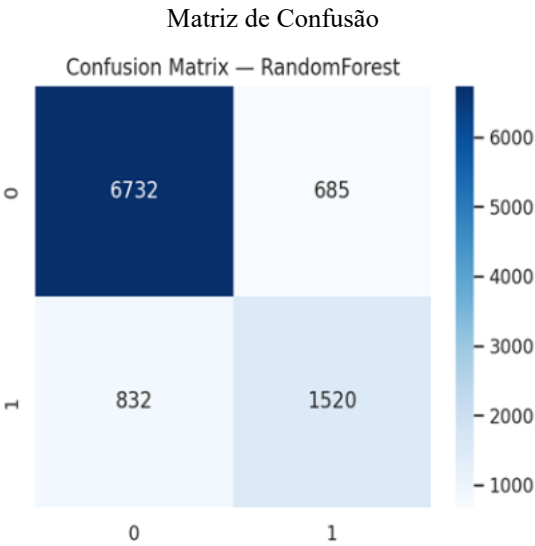


Fig. 2. Matriz de confusão de Random Forest, Arquivo: “data/confusion_rf.png”

O Random Forest apresenta melhor recall para a classe >50K quando comparado à Regressão Logística. O desempenho geral superior em acurácia reflete sua capacidade de capturar relações complexas entre variáveis.

	Comparação direta entre os Modelos		
	Critério	Logistic Regression	Random Forest
	Acurácia	0.8024	0.8447
	AUC-ROC	0.9007	0.8963
	Interpretabilidade	Excelente	Moderada
	Robustez linearidades	Baixa	Alta

D. Resumo

A Regressão Logística destaca-se pela interpretabilidade e alto AUC, sendo adequada quando transparência é crucial. O Random Forest obtém melhor acurácia e maior recall da classe minoritária, sendo mais adequado para aplicações práticas e cenários reais.

VI. DISCUSSÃO

Os experimentos demonstraram diferenças claras entre os dois modelos avaliados — Regressão Logística e Random Forest — tanto em desempenho quanto em comportamento estatístico. A escolha de ambos os modelos se mostrou adequada, pois eles representam abordagens distintas: um modelo linear probabilístico e um ensemble não linear baseado em árvores.

A **Regressão Logística** apresentou excelente capacidade discriminativa, refletida por um AUC-ROC de 0.9007, indicando que o modelo é altamente eficaz em separar as classes ao considerar probabilidades. Entretanto, sua acurácia de 0.8024 e o desempenho moderado na classe minoritária (>50K) revelam limitações derivadas da linearidade do modelo. Isso ocorre porque variáveis socioeconômicas possuem interações complexas que nem sempre podem ser capturadas por uma função logística com combinação linear de atributos.

Apesar dessas limitações, a interpretabilidade do modelo — especialmente via odds ratios — mostrou-se extremamente útil para compreender os fatores associados à renda elevada.

Por outro lado, o **Random Forest** obteve a maior acurácia entre os modelos avaliados (0.8447), demonstrando maior robustez e melhor capacidade de capturar relações não lineares. Entretanto, seu AUC-ROC ligeiramente inferior (0.8963) indica que, embora mais preciso na classificação, o modelo é marginalmente menos confiável na estimativa probabilística quando comparado à Regressão Logística. Ainda assim, sua habilidade de identificar variáveis importantes — como education-num, capital-gain-log e hours-per-week — complementa a análise interpretativa, mesmo que não ofereça a transparência estatística de um modelo linear.

A comparação entre os modelos evidencia um trade-off clássico:

- **Regressão Logística** → melhor explicabilidade, excelente AUC, porém limitada em padrões complexos.
- **Random Forest** → melhor acurácia, melhor recall da classe minoritária, maior robustez, porém menor interpretabilidade.

Essas observações mostram que ambos os modelos são adequados para o problema em questão, mas atendem a objetivos diferentes. Em cenários onde interpretabilidade é essencial (como políticas públicas ou análises sociológicas), a Regressão Logística é mais apropriada. Já em contextos de aplicação prática, onde a prioridade é desempenho, o Random Forest se mostra superior.

VII. CONCLUSÃO

Este trabalho apresentou a aplicação de dois modelos supervisionados — Regressão Logística e Random Forest — para prever se um indivíduo possui renda anual superior a USD 50K com base em atributos socioeconômicos. O estudo seguiu

uma metodologia rigorosa envolvendo pré-processamento, preparação de pipelines, divisão estratificada de treino e teste, e avaliação com métricas adequadas ao problema de classificação.

Os resultados mostraram que ambos os modelos apresentaram desempenho satisfatório:

- A **Regressão Logística** obteve AUC-ROC de **0.9007**, destacando-se como o modelo mais consistente na estimativa probabilística e o mais interpretável.
- O **Random Forest** alcançou a maior acurácia (**0.8447**), demonstrando capacidade superior de capturar padrões não lineares e relações complexas entre variáveis.

Conclui-se que ambos os modelos são adequados ao problema, cada um oferecendo vantagens distintas. A escolha depende da aplicação final: interpretabilidade e clareza estatística favorecem a Regressão Logística, enquanto precisão e robustez favorecem o Random Forest.

Como trabalhos futuros, sugere-se:

1. Avaliar técnicas de balanceamento (SMOTE, undersampling) para melhorar o recall da classe >50K.
2. Investigar regularização na Regressão Logística (L1/L2) para melhor estabilidade.
3. Testar ensemble híbrido (Logistic + Random Forest) ou modelos mais modernos (LightGBM, CatBoost) caso o escopo permita.
4. Realizar análise de fairness e viés, dado o caráter sensível do dataset.

Essas extensões podem aprofundar o entendimento do problema e aprimorar a qualidade das previsões em aplicações reais.

VIII. REFERÊNCIAS

- [1] D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant, Applied Logistic Regression, 3rd ed. Wiley, 2013.
- [2] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [3] Kaggle, "Salary Prediction Classification Dataset."