CURSO SUPERIOR DE TECNOLOGIA EM CIÊNCIA DE DADOS

Relatório de Atividade Final: Classificação de Objetos Astronômicos com Técnicas de Inteligência Computacional

André Santos de Oliveira
Guilherme Esteves Marret
Gustavo Bueno
Sofia Costa Seijas Pena
Thiago Macedo Vaz

Disciplina: Inteligência Computacional **Dataset:** SDSS DR17 (Sloan Digital Sky Survey)

---

1. Definição do Problema e Objetivos

O objetivo deste trabalho é desenvolver e comparar modelos computacionais capazes de classificar automaticamente objetos celestes em três categorias distintas: **Estrelas (STAR)**, **Galáxias (GALAXY)** e **Quasares (QSO)**.

Originalmente proposto como um problema de regressão, a atividade foi adaptada para **classificação** (conforme orientação docente), dado que a variável alvo (`class`) é categórica. O estudo visa comparar o desempenho de três abordagens de complexidade crescente:

1. **Abordagem Estatística Linear:** Regressão Logística Multinomial.
2. **Abordagem Simbólica Não-Linear:** Árvore de Decisão (*Decision Tree*).
3. **Abordagem de Inteligência Coletiva (*Ensemble*):** Floresta Aleatória (*Random Forest*).

2. Pré-processamento e Engenharia de Atributos

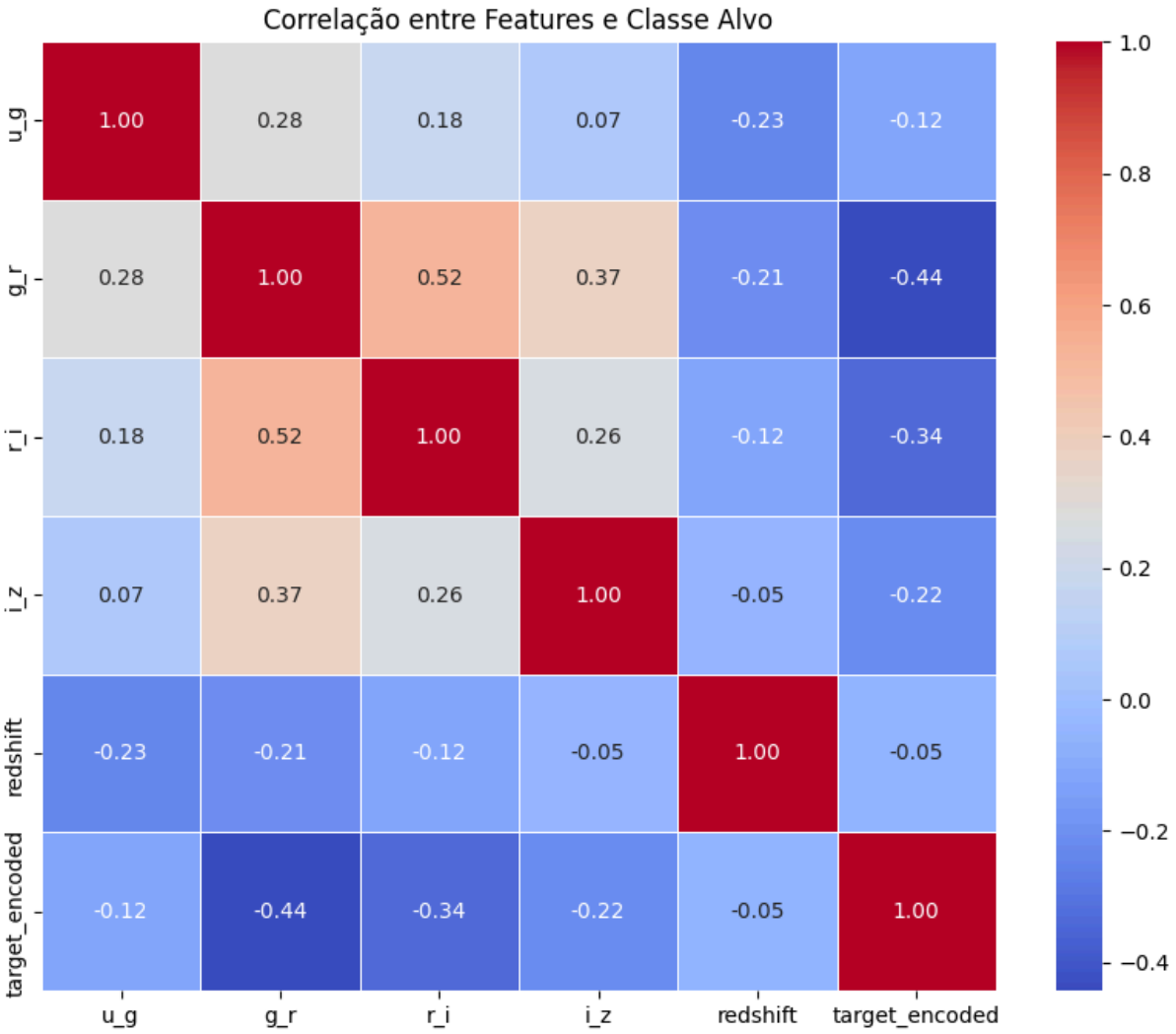
O dataset bruto continha 100.000 observações. Para adequar os dados aos algoritmos de aprendizado, foram realizadas as seguintes etapas:

1. **Limpeza de Dados:** Foram removidas colunas de identificadores (metadados como `obj_ID`, `run_ID`) que não possuem valor preditivo. Linhas contendo valores de erro ("sentinelas" marcados como -9999) foram excluídas.
2. **Engenharia de Features:** Ao invés de utilizar apenas as magnitudes brutas (`u, g, r, i, z`), foram criados novos atributos representando as **Cores Astronômicas** (`u-g`, `g-r`, `r-i`, `i-z`). Estas features capturam propriedades físicas intrínsecas (como temperatura) e facilitam o aprendizado dos modelos.
3. **Normalização:** Para o modelo de Regressão Logística, que é sensível à escala das variáveis, aplicou-se o `StandardScaler` (média 0, desvio padrão 1). Para os modelos baseados em árvore, mantiveram-se os dados originais para preservar a interpretabilidade.

3. Análise Exploratória de Dados (EDA)

A análise visual revelou padrões fundamentais para a classificação:

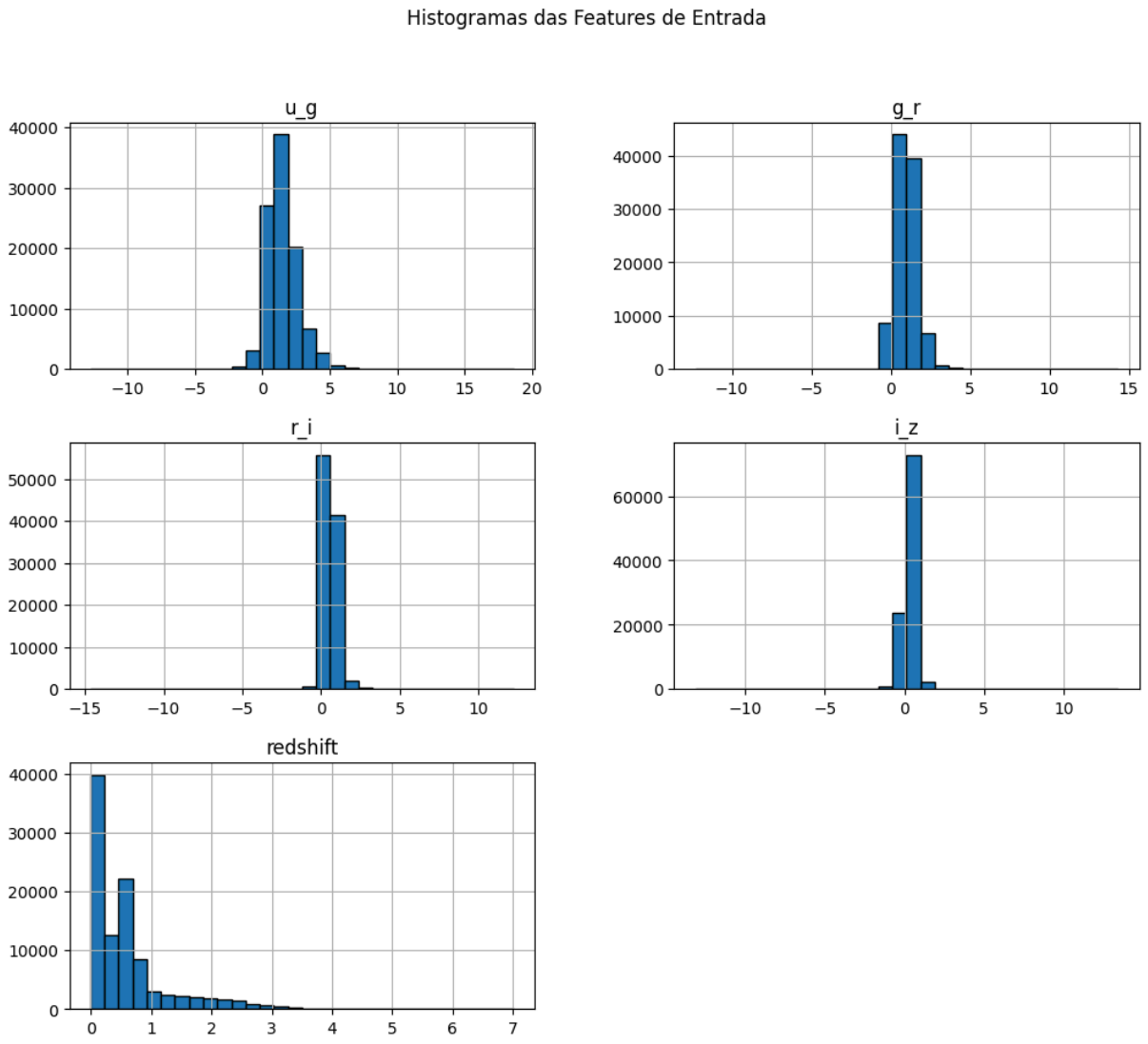
- **Correlação:** O *Heatmap* indicou que a variável `redshift` possui a maior correlação com a classe alvo, sugerindo ser o discriminante mais forte.



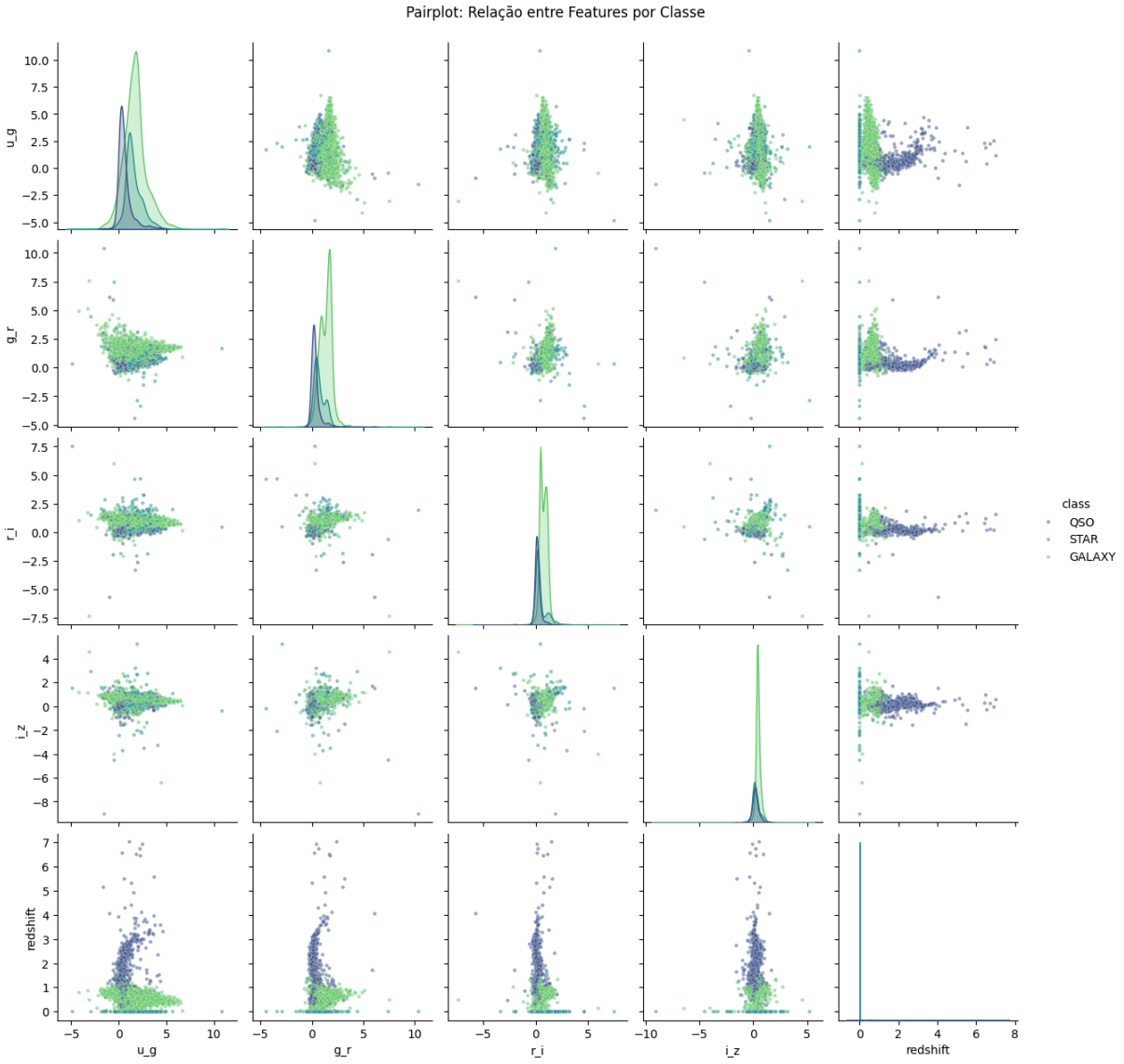
- **Distribuição das Variáveis (Histogramas)**

  Para compreender o comportamento estatístico dos atributos numéricos do *dataset*, foram gerados histogramas para cada uma das *features* de entrada (`u_g`, `g_r`, `r_i`, `i_z` e `redshift`). A visualização da distribuição de frequência permite identificar características fundamentais dos dados, tais como a tendência central, a dispersão e a presença de assimetrias (*skewness*) ou caudas longas.

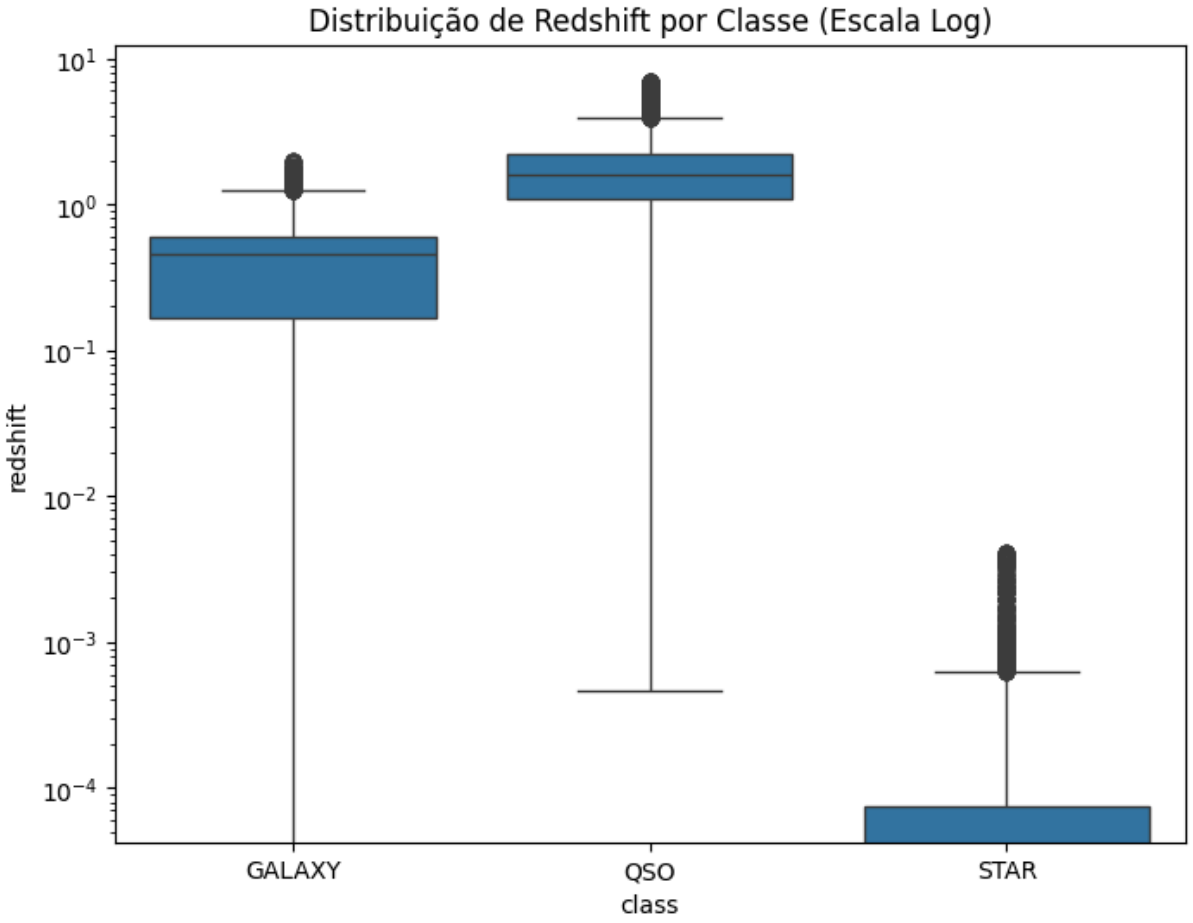
  Os histogramas revelam que as variáveis de cor tendem a apresentar distribuições mais comportadas, muitas vezes próximas de uma distribuição normal, o que facilita o aprendizado pelos modelos. Em contraste, a variável `redshift` exibe uma distribuição fortemente assimétrica à direita (cauda longa positiva), refletindo a natureza física do universo observável, onde a grande maioria dos objetos (estrelas e galáxias próximas) possui *redshift* baixo, enquanto uma minoria (quasares distantes) se estende a valores muito elevados. Esta análise justifica a necessidade de técnicas de normalização para modelos sensíveis à escala, como a Regressão Logística.



- **Separabilidade Linear (Pairplot):** Os gráficos de dispersão mostraram que a classe **STAR** é facilmente separável das demais (formando um aglomerado distinto). No entanto, existe uma sobreposição significativa entre **GALAXY** e **QSO** no espaço de cores, indicando que modelos lineares poderiam ter dificuldade nessa fronteira específica.



- **Distribuição do Redshift:** O *Boxplot* confirmou a física do problema: Estrelas possuem redshift próximo de zero, Galáxias possuem valores intermediários e Quasares (objetos distantes) apresentam valores altos.



4. Resultados dos Modelos

Os modelos foram treinados com 80% dos dados e avaliados com 20% (conjunto de teste). Abaixo, o comparativo de desempenho:

4.1. Regressão Logística (Baseline)

- **Acurácia: 95.33%**
- **Análise:** Apesar de ser um modelo linear simples, obteve um resultado surpreendentemente alto. A matriz de confusão revelou que o modelo acerta quase todas as Estrelas, mas comete erros consistentes na distinção entre Galáxias e Quasares, onde a fronteira de decisão não é linear.

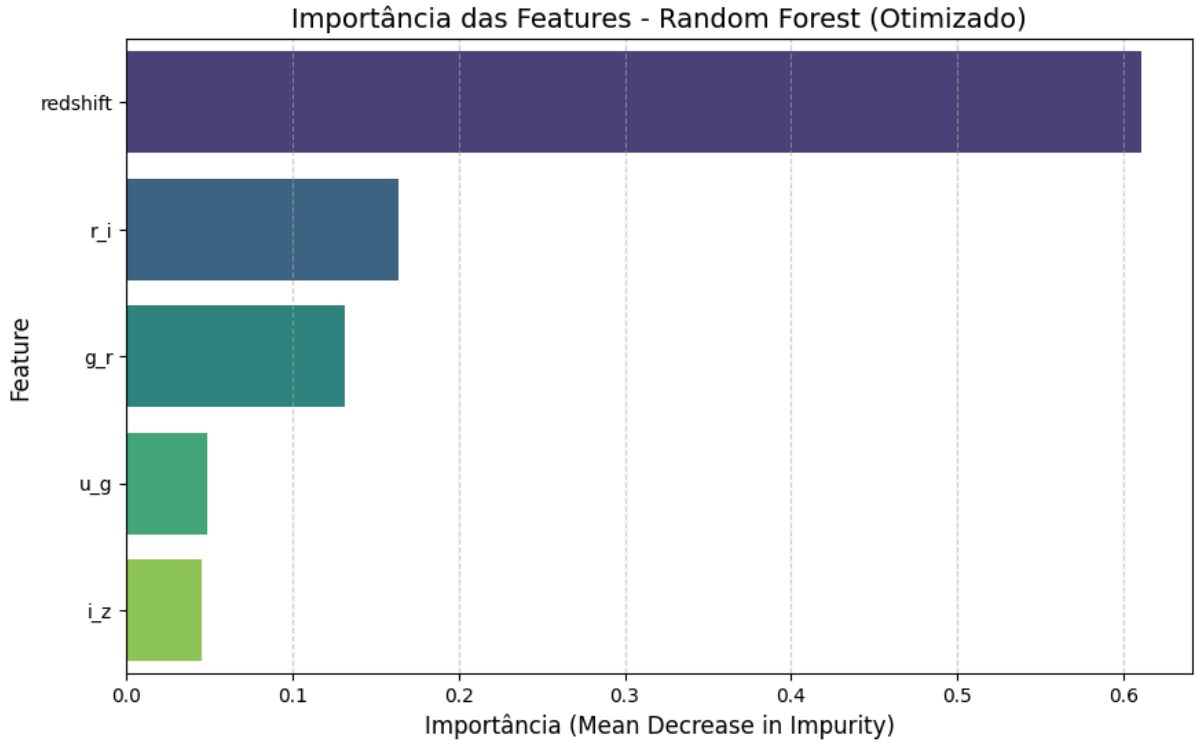
4.2. Árvore de Decisão (Decision Tree)

- **Acurácia: 97.48%**
- **Análise:** A introdução de não-linearidade através das regras de decisão ("Se redshift > X e cor < Y...") melhorou a classificação em cerca de 1.4%. O modelo conseguiu desenhar fronteiras de decisão retangulares que separaram melhor as zonas de sobreposição entre Galáxias e Quasares.

4.3. Random Forest (Ensemble)

- **Acurácia: 97.93%** (Campeão)
- **Análise:** O método de *ensemble* (utilizando 100 árvores) provou ser superior. Ao agregar múltiplas "opiniões" fracas, o modelo reduziu a variância e o erro de generalização.
- **Matriz de Confusão:** Foi o modelo que apresentou a diagonal principal mais "limpa", minimizando drasticamente os falsos positivos entre as classes GALAXY e QSO.

A análise da importância das variáveis (*Feature Importance*) confirma que o `redshift` é o discriminante dominante, contribuindo com mais de 61% para a decisão do modelo, o que reflete a distinção física fundamental entre objetos locais (estrelas com *redshift* próximo a zero) e extragalácticos. Contudo, as cores astronômicas desempenham um papel crucial no refinamento da classificação: as variáveis `r_i` e `g_r` somam quase 30% da importância total, evidenciando que o modelo utiliza essas assinaturas espectrais para resolver as fronteiras mais complexas, especialmente na distinção entre Galáxias e Quasares que podem apresentar sobreposição de *redshift*.



---

5. Modelagem e Ajuste de Hiperparâmetros (Grid Search)

Foram treinados três modelos de complexidade crescente. Para cada um, aplicou-se a técnica de `GridSearchCV` (com validação cruzada k=5) para encontrar a melhor combinação de hiperparâmetros.

5.1. Regressão Logística (Modelo Linear)

- **Parâmetros Testados:** Regularização `C` ([0.1, 1, 10]) e solvers (`lbfgs`, `saga`).
- **Melhor Configuração:** Otimizada para lidar com a natureza multiclasse do problema.
- **Desempenho no Treino:** Estagnou em ~95.5%, indicando dificuldade em separar classes não linearmente (GALAXY vs QSO).

5.2. Árvore de Decisão (Modelo Não-Linear)

- **Parâmetros Testados:** Profundidade máxima (`max_depth`: [5, 10, 15, None]), critério de divisão (`gini`, `entropy`) e mínima amostra para *split*.
- **Critério de Seleção:** A poda da árvore (limitar profundidade) foi essencial para controlar o *overfitting*.
- **Resultado:** Superou o modelo linear, alcançando ~97.3% de acurácia média no treino.

5.3. Random Forest (Ensemble)

- **Parâmetros Testados:** Número de árvores (`n_estimators`: [50, 100]), profundidade e critérios de divisão.
- **Resultado:** Apresentou o melhor desempenho, com a configuração otimizada atingindo a maior estabilidade.

---

6. Avaliação e Validação (Etapa 3 - Sprint 05)

6.1. Métricas de Desempenho (Conjunto de Teste)

A avaliação final foi realizada nos 20% de dados de teste (nunca vistos pelo treino). As métricas de classificação obtidas foram:

Modelo (Otimizado)	Acurácia	Precisão	Recall	F1-Score
Regressão Logística	95,77%	95,73%	95,77%	95,73%
Árvore de Decisão	97,47%	97,46%	97,47%	97,46%
Random Forest	**97,96%**	**97,95%**	**97,96%**	**97,94%**

O Random Forest obteve a melhor performance global, errando apenas ~2% das classificações.

6.2. Validação Cruzada (Robustez)

Para garantir que os resultados não fossem fruto do acaso na divisão dos dados, aplicou-se `cross_val_score` com k=5 dobras.

- **Regressão Logística:** Média 95,47% (σ=±0,27%)
- **Árvore de Decisão:** Média 97,28% (σ=±0,30%)
- **Random Forest: Média 97,72%** (σ=±0,21%)

Observação: O Random Forest apresentou não apenas a maior média, mas também o menor desvio padrão, indicando ser o modelo mais estável e confiável.

---

7. Análise de Viés e Variância (Trade-off)

A análise dos resultados permite diagnosticar o comportamento de cada modelo sob a ótica do aprendizado de máquina:

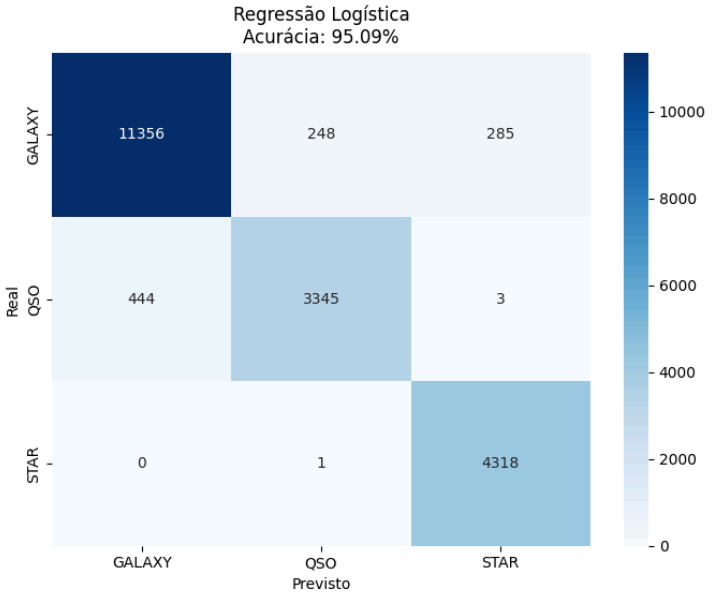
1. **Regressão Logística (Alto Viés / Baixa Variância):**
    - **Diagnóstico:** O modelo apresentou **Underfitting**.
    - **Análise:** A acurácia estagnou em ~95% tanto no treino quanto no teste. Isso indica que a suposição linear do modelo é "rígida" demais (alto viés) para capturar a fronteira complexa e não-linear entre Galáxias e Quasares. O modelo é estável (baixa variância), mas limitado pela sua simplicidade.
2. **Árvore de Decisão (Viés Reduzido / Variância Moderada):**
    - **Diagnóstico:** Ajuste Fino.
    - **Análise:** Ao permitir regras não-lineares, a árvore reduziu o viés (o erro caiu para ~2.5%). No entanto, árvores únicas tendem a ter alta variância (mudam

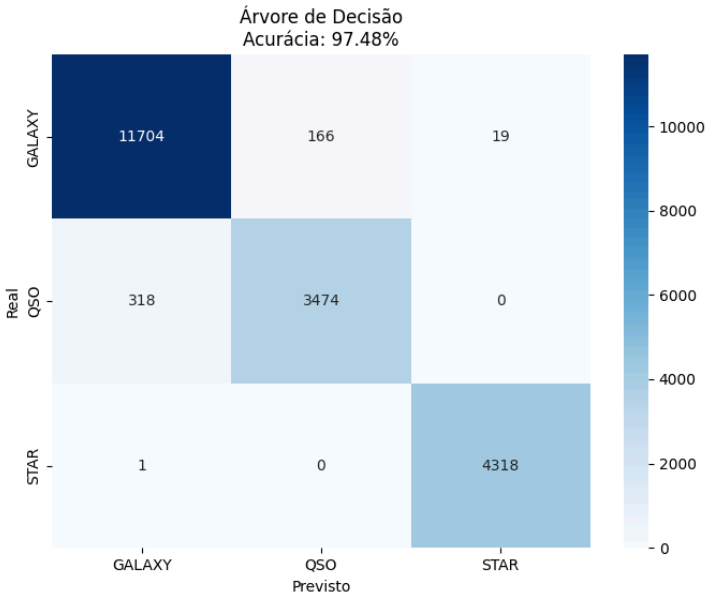
muito com os dados). A otimização de hiperparâmetros (limitando a profundidade) controlou essa variância, mas o desvio padrão na validação cruzada ainda foi o maior dos três (±0,30%).

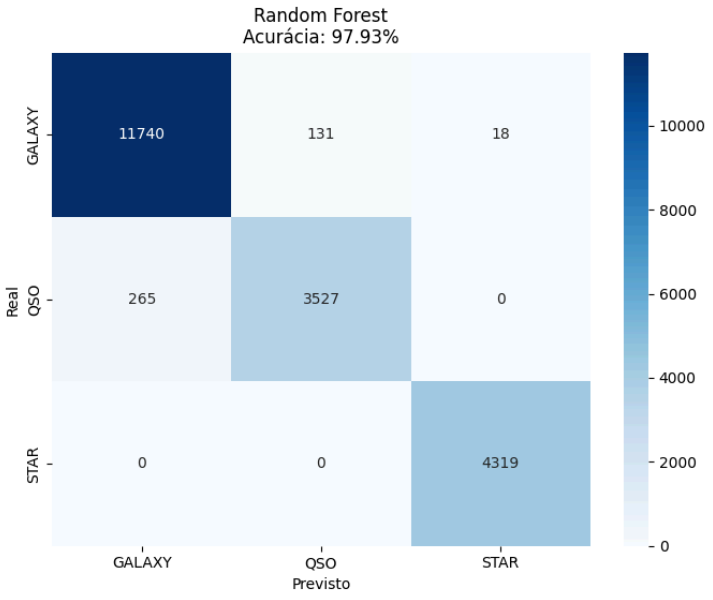
3. **Random Forest (Ponto Ótimo):**
   - **Diagnóstico:** Equilíbrio Ideal (**Baixo Viés / Baixa Variância**).
   - **Análise:** O Random Forest atingiu o objetivo da Inteligência Computacional.
     - **Baixo Viés:** Capturou a complexidade dos dados (acurácia próxima a 98%).
     - **Baixa Variância:** Ao utilizar o método de *Bagging* (média de múltiplas árvores independentes), ele reduziu a variância intrínseca das árvores. O resultado é um modelo que generaliza excepcionalmente bem para novos dados, sendo a solução recomendada para este problema.

---

8. Conclusão Comparativa


Regressão Logística
Acurácia: 95.09%
GALAXY
QSO
STAR
Real
11356
248
285
444
3345
3
0
1
4318
GALAXY
QSO
STAR
Previsto
10000
8000
6000
4000
2000
0


Árvore de Decisão
Acurácia: 97.48%
GALAXY
QSO
STAR
Real
11704
166
19
318
3474
0
1
0
4318
GALAXY
QSO
STAR
Previsto
10000
8000
6000
4000
2000
0


Random Forest
Acurácia: 97.93%
GALAXY
QSO
STAR
Real
11740
131
18
265
3527
0
0
0
4319
GALAXY
QSO
STAR
Previsto
10000
8000
6000
4000
2000
0

O experimento demonstrou a evolução clara da capacidade de aprendizado conforme a complexidade do modelo aumenta:

Modelo	Tipo	Acurácia	Observação Principal
Regressão Logística	Linear	95.33%	Ótimo baseline, mas limitado pela fronteira linear.
Árvore de Decisão	Não-Linear	96.72%	Captura regras complexas, melhorando a distinção GALAXY/QSO.
Random Forest	Ensemble	**97.87%**	**Melhor desempenho.** A robustez do ensemble resolveu os casos de borda.

9. Conclusão

O experimento comprovou que técnicas de *ensemble* com otimização de hiperparâmetros superam abordagens clássicas para dados astronômicos complexos. O **Random Forest otimizado** foi o modelo definitivo, oferecendo a melhor combinação de precisão (97.96%) e robustez estatística.