



Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital



RELATÓRIO DO PROJETO 1: Cursos de Graduação no Brasil - Classificação e Regressão

Francisco Willem Romão Moreira
20220036966

Natal-RN
2024

RESUMO

Este relatório mostra um projeto desenvolvido na disciplina de Análise Computacional da Aprendizagem, onde o autor criou modelos de aprendizado de máquina, especialmente classificação e regressão com uma base dados extraída do portal de dados abertos do Ministério da Educação (MEC) sobre os cursos de graduação no Brasil. Além disso, é mostrado as etapas de pré-processamento, incluindo limpeza, balanceamento e transformação. Por fim, é exposto os resultados das métricas dos modelos, sendo 79% de acurácia para o modelo de classificação e um R^2 de 0.64 para o modelo de regressão.

Palavras-chave: classificação, regressão, cursos de graduação Brasil.

LISTA DE FIGURAS

1	Quantidade de vagas autorizadas por curso antes da remoção de outliers. .	5
2	Quantidade de vagas autorizadas por curso depois da remoção de outliers.	6
3	Carga horária por curso antes da remoção de outliers.	6
4	Carga horária por curso depois da remoção de outliers.	6
5	Diagrama do modelo de classificação no orange.	7
6	Seleção de variáveis para o modelo de classificação.	7
7	Discretização.	8
8	Configuração da árvore.	8
9	Visualização de apenas três níveis da árvore.	9
10	Configuração dos parâmetros da random forest.	9
11	Diagrama do modelo de regressão no orange.	9
12	Verificação de dados nulos.	10
13	Avaliação geral do modelo de classificação.	11
14	Matriz de confusão para Árvore de Decisão.	12
15	Matriz de confusão para Florestas Aleatórias.	12
16	Avaliação geral do modelo de regressão.	14
17	Correlação de Pearson.	15
18	Gráfico de regressão GRAU por CARGA HORARIA.	16

LISTA DE TABELAS

SUMÁRIO

1	INTRODUÇÃO	5
1.1	Problema.....	5
1.2	Base de dados	5
2	METODOLOGIA	5
2.1	Pré-processamento: remoção de outliers	5
2.2	Modelo de Classificação.....	7
2.2.1	<i>Balanceamento</i>	7
2.2.2	<i>Seleção de variáveis</i>	7
2.2.3	<i>Discretização</i>	7
2.2.4	<i>Escolha do algoritmo</i>	8
2.3	Modelo de Regressão	9
2.3.1	<i>Seleção de variáveis</i>	9
2.3.2	<i>Verificação de dados nulos</i>	10
2.3.3	<i>Pré-processamento: Normalização</i>	10
2.3.4	<i>Pré-processamento: Transformação</i>	10
2.3.5	<i>Escolha do algoritmo</i>	11
3	RESULTADOS	11
3.1	Modelo de Classificação.....	11
3.2	Modelo de Regressão	12
4	CONCLUSÃO.....	16
	REFERÊNCIAS.....	17

1 INTRODUÇÃO

1.1 Problema

Inicialmente, o problema foi focado na classificação do risco de extinção de um curso de graduação, utilizando a variável `SITUACAO_CURSO` da base de dados, que possui três categorias: *Em atividade*, *Em extinção* e *Extinto*.

Em seguida o foco foi criar um modelo de regressão para prever qual a carga horária ideal para um curso com base em características como categoria administrativa, organização acadêmica, grau, modalidade, situação, quantidade de vagas e região.

1.2 Base de dados

Os dados advêm da página de dados abertos do Ministério da Educação (MEC, 2022). O dataset foi coletado diretamente dessa página por meio do download do arquivo csv. As principais características do dataset incluem os cursos de graduação (Licenciatura, Bacharelado, Tecnológico, Sequencial e ABI - Área Básica de Ingresso) no Brasil por: código da Instituição de Educação Superior (IES); nome da IES; categoria da IES; organização acadêmica; código do curso; nome do curso; grau; área OCDE; modalidade de ensino (presencial ou EaD); situação do curso (ativo ou inativo); vagas autorizadas; carga horária; segmentadas por código do município (IBGE); município; UF; região.

2 METODOLOGIA

2.1 Pré-processamento: remoção de outliers

A primeira etapa foi explorar a base, com isso foi descoberto que as colunas com valores numéricos (`QT_VAGAS_AUTORIZADA` e `CARGA_HORARIA`) possuíam outliers, conforme as figuras abaixo.

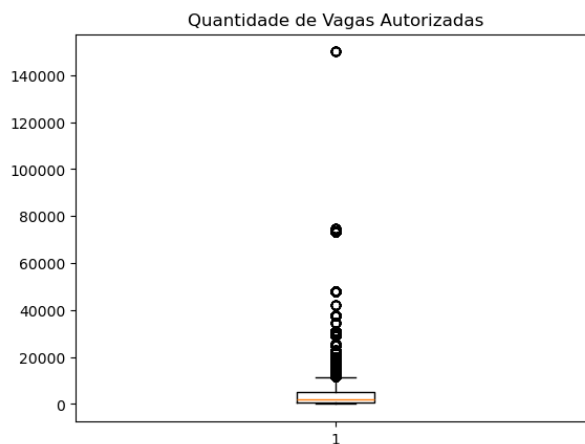


Figura 1 – Quantidade de vagas autorizadas por curso antes da remoção de outliers.

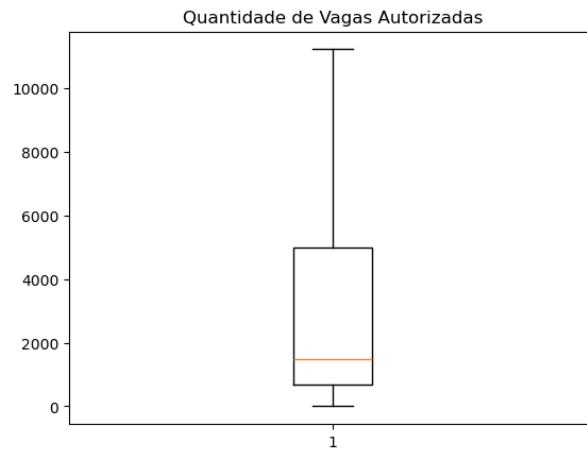


Figura 2 – Quantidade de vagas autorizadas por curso depois da remoção de outliers.

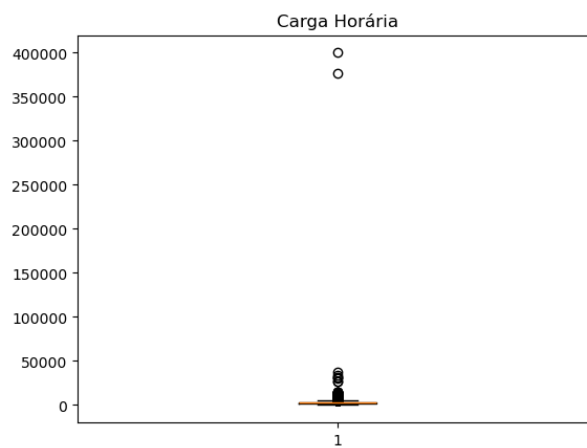


Figura 3 – Carga horária por curso antes da remoção de outliers.

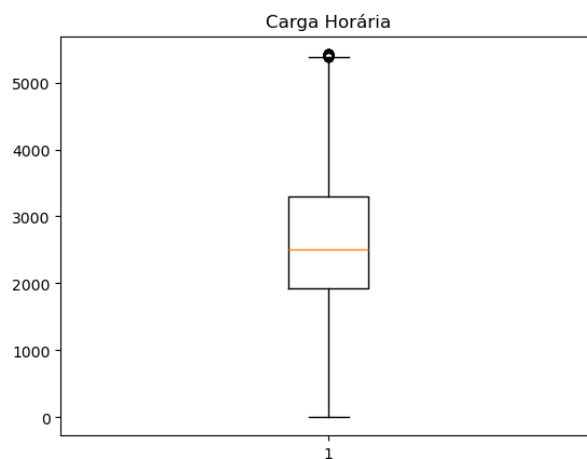


Figura 4 – Carga horária por curso depois da remoção de outliers.

Para tratar esses outliers foi aplicado o método Interquartile Range (IQR). O IQR é uma medida de dispersão que representa a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) de um conjunto de dados, abrangendo os valores centrais dos 50% intermediários.

Com essa remoção de outliers, o dataset passou de 902.676 para 825.321 linhas, o que representa pouco mais de 8% de redução.

2.2 Modelo de Classificação

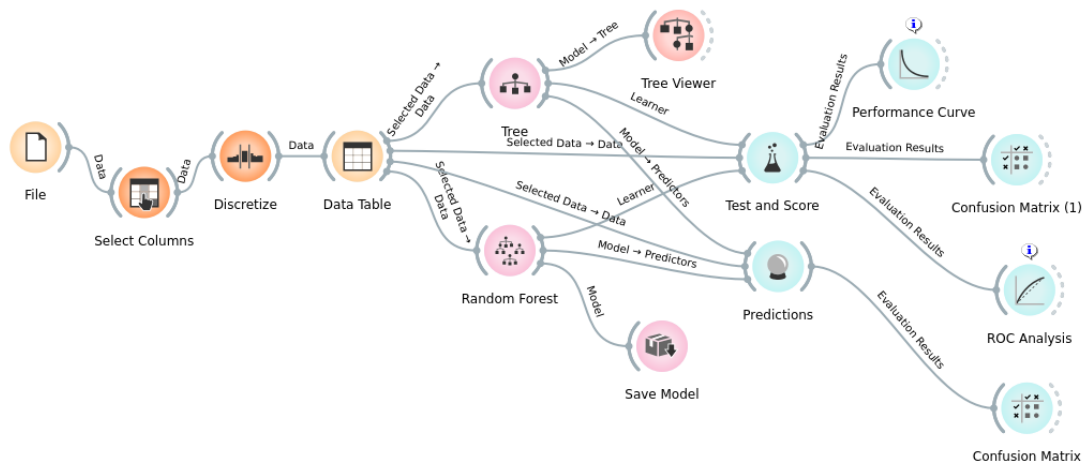


Figura 5 – Diagrama do modelo de classificação no Orange.

2.2.1 Balanceamento

A primeira etapa a ser feita para o modelo de classificação foi o balanceamento dos dados, que consiste em equilibrar os valores das classes majoritárias. Por exemplo, o dataset após a limpeza ficou com 776.097 cursos. Em atividade, 39.017 cursos Extintos e 10.207 Em extinção, sendo assim, o balanceamento selecionou aleatoriamente cursos buscando equilibrar as classes para que cada uma delas tivesse 10.207 cursos. Assim o dataset final balanceado ficou com 30.621 cursos.

2.2.2 Seleção de variáveis

A segunda etapa consistiu em selecionar as variáveis que seriam utilizadas para o modelo de classificação, como também o alvo. Chegando na seguinte configuração:

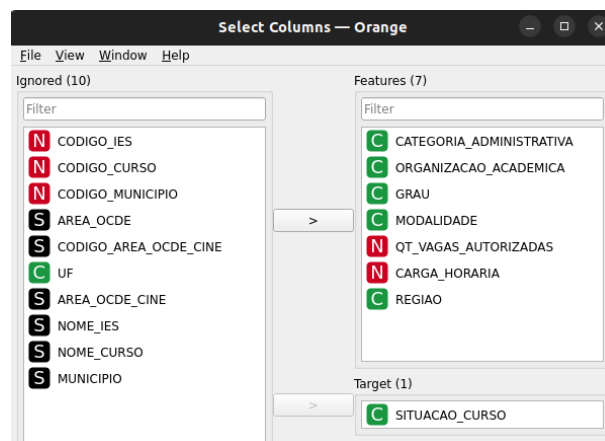


Figura 6 – Seleção de variáveis para o modelo de classificação.

2.2.3 Discretização

Em seguida foi feita uma discretização nas colunas numéricas de quantidade de vagas e carga horária. A discretização é o processo de transformar dados contínuos em

dados categóricos, dividindo o intervalo de valores contínuos em um número finito de intervalos ou categorias.

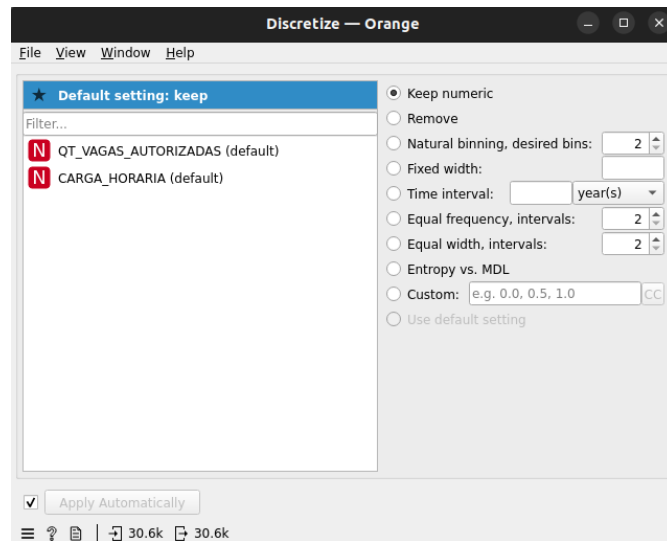


Figura 7 – Discretização.

2.2.4 Escolha do algoritmo

O principal algoritmo utilizado foi o de Árvore de Decisão (Decision Tree). O algoritmo de árvore de decisão é um método de aprendizado supervisionado que utiliza um modelo em forma de árvore para tomar decisões, dividindo iterativamente os dados em subconjuntos com base em características que fornecem a maior distinção entre as classes até que as folhas representem previsões ou valores finais.

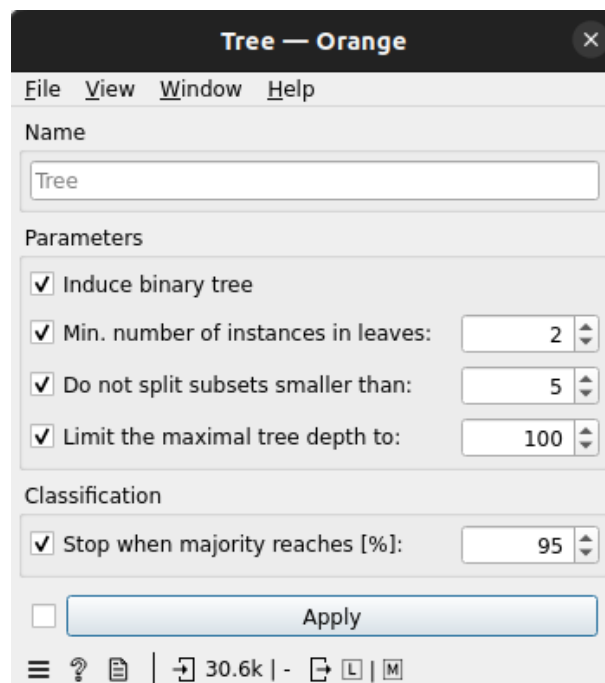


Figura 8 – Configuração da árvore.

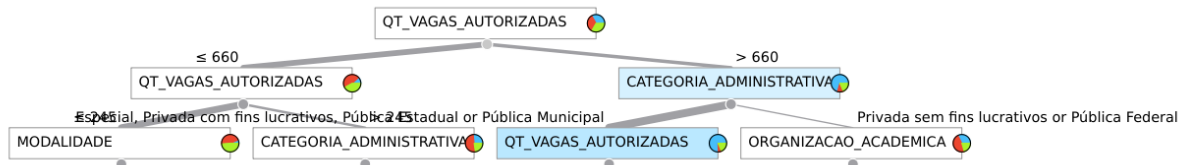


Figura 9 – Visualização de apenas três níveis da árvore.

Também decidi utilizar o algoritmo de Florestas Aleatórias (Random Forest). O Random Forest é um método que cria múltiplas árvores de decisão independentes durante o treinamento e combina suas previsões para melhorar a precisão do modelo.

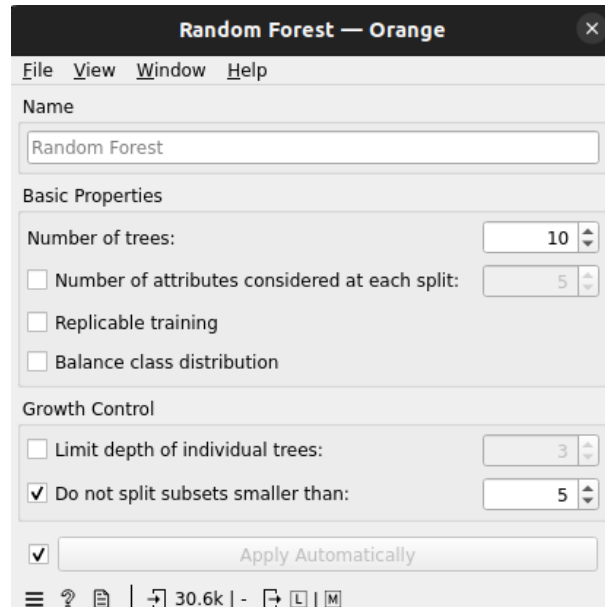


Figura 10 – Configuração dos parâmetros da random forest.

2.3 Modelo de Regressão

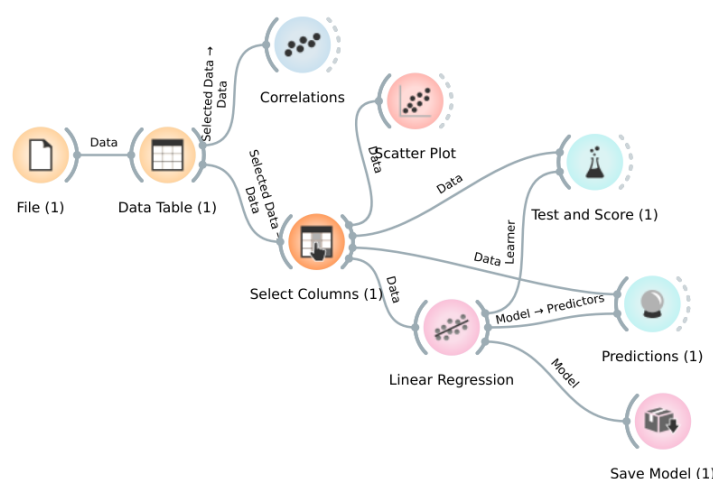


Figura 11 – Diagrama do modelo de regressão no orange.

2.3.1 Seleção de variáveis

Comecei selecionando algumas variáveis:

- CATEGORIA_ADMINISTRATIVA
- ORGANIZACAO_ACADEMICA
- GRAU
- MODALIDADE
- SITUACAO_CURSO
- QT_VAGAS_AUTORIZADAS
- CARGA_HORARIA
- REGIAO

2.3.2 Verificação de dados nulos

A primeira etapa foi verificar os dados nulos para não dar problema no algoritmo de regressão.

```
df.isnull().sum()

CATEGORIA_ADMINISTRATIVA    0
ORGANIZACAO_ACADEMICA       0
GRAU                        0
MODALIDADE                   0
SITUACAO_CURSO              0
QT_VAGAS_AUTORIZADAS       0
CARGA_HORARIA               0
REGIAO                      0
dtype: int64
```

Figura 12 – Verificação de dados nulos.

A qualidade do dataset se mostra boa, pois não foi preciso fazer o tratamento dos valores nulos.

2.3.3 Pré-processamento: Normalização

Para o algoritmo de regressão performar bem, foi necessário normalizar as variáveis numéricas QT_VAGAS_AUTORIZADAS e CARGA_HORARIA, para isso, usei o método StandardScaler da biblioteca Scikit-Learn. O método de normalização StandardScaler transforma os dados para que tenham média zero e desvio padrão igual a um, ajustando os valores de cada característica para que fiquem na mesma escala e melhorem o desempenho dos algoritmos de aprendizado de máquina.

2.3.4 Pré-processamento: Transformação

Também foi necessário fazer uma transformação das variáveis categóricas, para isso usei o método OneHotEncoder também do Scikit-Learn. O método OneHotEncoder converte variáveis categóricas em uma representação binária, criando colunas adicionais para cada categoria possível, onde o valor 1 indica a presença da categoria e 0 a ausência, permitindo que algoritmos de aprendizado de máquina que de antemão só funcionam com variáveis numéricas, trabalhem com dados categóricos.

2.3.5 Escolha do algoritmo

Para o modelo de regressão usei a regressão linear. A regressão linear é um método estatístico que modela a relação entre uma variável dependente e uma ou mais variáveis independentes, ajustando uma linha reta (ou hiperplano, no caso de múltiplas variáveis) que minimiza a soma dos quadrados dos erros entre as previsões e os valores reais.

3 RESULTADOS

3.1 Modelo de Classificação

A seguir mostro resultados gerais acerca do modelo de classificação. Para entender melhor os resultados da Figura [13], precisamos explicar algumas métricas.

A **Curva de ROC** (AUC) refere-se à área sob a curva ROC (Receiver Operating Characteristic). AUC é uma métrica de desempenho para modelos de classificação binária que mede a capacidade do modelo de distinguir entre classes positivas e negativas.

A **Acurácia** (CA) refere-se a proporção de previsões corretas (tanto verdadeiros positivos quanto verdadeiros negativos) em relação ao total de exemplos.

O **F1-Score** (F1) é média harmônica da precisão e da revocação. É útil quando é necessário um equilíbrio entre precisão e revocação.

A **Precisão** (Prec) é proporção de verdadeiros positivos em relação ao total de previsões positivas. Indica a qualidade das previsões positivas do modelo.

Já a **Revocação** (Recall) é a proporção de verdadeiros positivos em relação ao total de exemplos reais positivos. Mede a capacidade do modelo de encontrar todos os exemplos positivos.

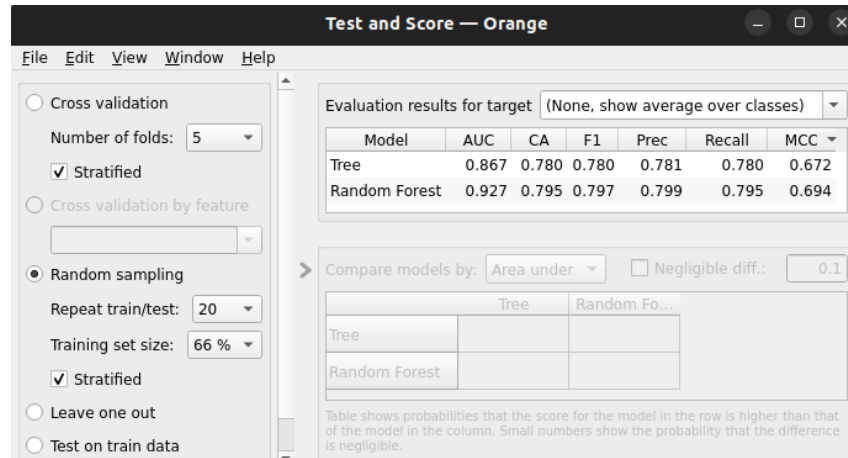


Figura 13 – Avaliação geral do modelo de classificação.

Também temos outra métrica de avaliação muito importante, a Matriz de Confusão. A matriz de confusão é uma tabela usada para avaliar o desempenho de um modelo de classificação, mostrando as previsões corretas e incorretas em cada categoria. Aqui está uma explicação simples:

- **Verdadeiros Positivos (TP):** Casos em que o modelo previu corretamente a classe positiva.
- **Verdadeiros Negativos (TN):** Casos em que o modelo previu corretamente a classe negativa.

- **Falsos Positivos (FP):** Casos em que o modelo previu incorretamente a classe positiva (falsos alarmes).
- **Falsos Negativos (FN):** Casos em que o modelo previu incorretamente a classe negativa (erros de omissão).

A estrutura da matriz de confusão é geralmente assim:

	Previsto Positivo	Previsto Negativo
Real Positivo	TP	FN
Real Negativo	FP	TN

Ela ajuda a entender onde o modelo está errando e é especialmente útil para detectar desequilíbrios nas classificações. Uma matriz de confusão é considerada boa quando os valores na diagonal principal (verdadeiros positivos e verdadeiros negativos) são altos e os valores fora da diagonal principal (falsos positivos e falsos negativos) são baixos. Nas Figuras [14] e [15] temos os valores em proporção na unidade de porcentagem para fins didáticos.

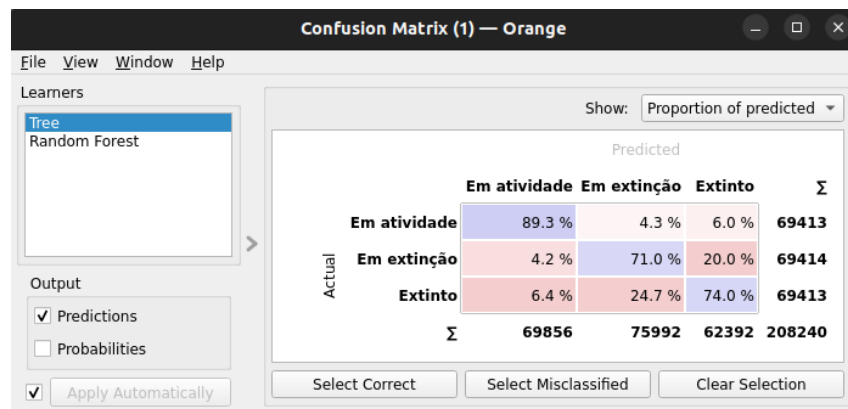


Figura 14 – Matriz de confusão para Árvore de Decisão.

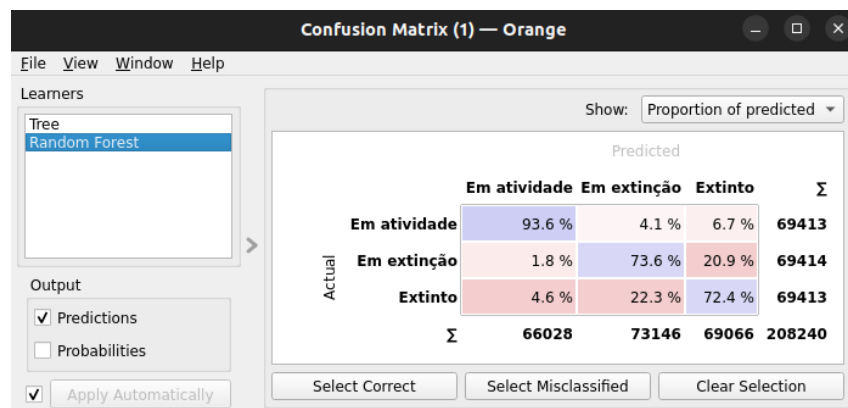


Figura 15 – Matriz de confusão para Florestas Aleatórias.

3.2 Modelo de Regressão

Por fim, os resultados do modelo de Regressão. Aqui estão explicações breves de MSE, RMSE, MAPE e R^2 , que são métricas comuns para avaliar modelos de regressão:

- **MSE (Mean Squared Error):**

- É a média dos quadrados dos erros, ou seja, a diferença entre os valores previstos pelo modelo e os valores reais.
- Fórmula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- MSE penaliza erros grandes mais fortemente, pois os erros são elevados ao quadrado.

- **RMSE (Root Mean Squared Error):**

- É a raiz quadrada da média dos quadrados dos erros. Fornece uma medida do erro em unidades da variável de resposta.
- Fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- RMSE é mais interpretável do que MSE porque está na mesma escala que os dados.

- **MAPE (Mean Absolute Percentage Error):**

- É a média das diferenças absolutas entre os valores previstos e os valores reais, expressas como uma porcentagem dos valores reais.
- Fórmula:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- MAPE é útil para entender o erro em termos relativos, mas pode ser problemático se os valores reais forem muito próximos de zero.

- **R^2 (Coeficiente de Determinação):**

- Mede a proporção da variância total dos dados que é explicada pelo modelo. Um valor de R^2 próximo de 1 indica que o modelo explica bem a variabilidade dos dados.
- Fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

onde \bar{y} é a média dos valores reais.

- R^2 varia de 0 a 1, onde 0 indica que o modelo não explica nenhuma variabilidade e 1 indica que explica toda a variabilidade dos dados.

Essas métricas ajudam a avaliar a precisão e a eficácia dos modelos de regressão, cada uma fornecendo uma perspectiva diferente sobre o desempenho do modelo.

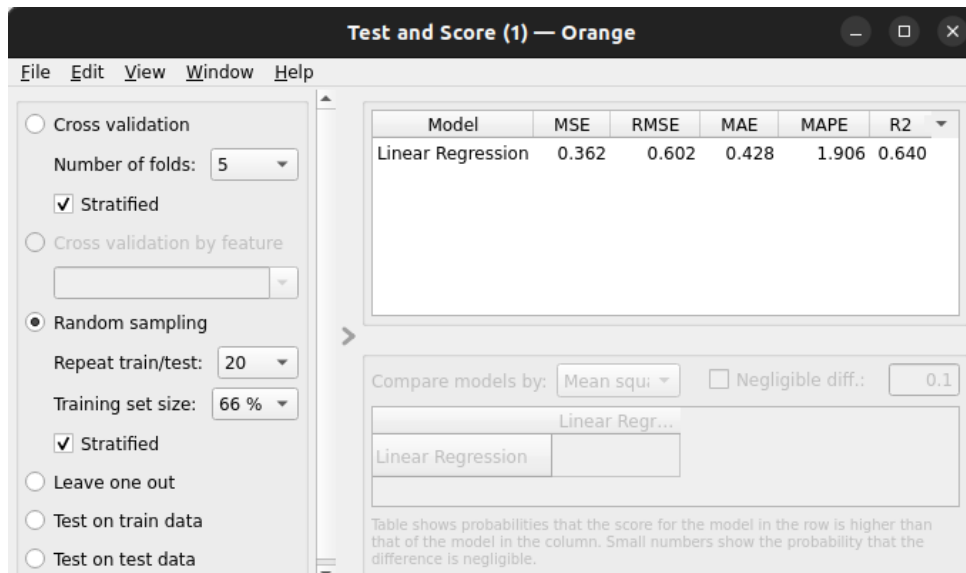


Figura 16 – Avaliação geral do modelo de regressão.

Para construção desse modelo também foi usada a correlação de Pearson, conhecida como coeficiente de correlação de Pearson, mede a força e a direção da relação linear entre duas variáveis quantitativas.

- **Valor:** O coeficiente de correlação de Pearson varia de -1 a 1.
 - **1** indica uma correlação positiva perfeita (quando uma variável aumenta, a outra também aumenta de forma linear).
 - **-1** indica uma correlação negativa perfeita (quando uma variável aumenta, a outra diminui de forma linear).
 - **0** indica nenhuma correlação linear (as variáveis não têm uma relação linear discernível).

- **Fórmula:**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Onde x_i e y_i são os valores das variáveis, e \bar{x} e \bar{y} são suas médias.

- **Interpretação:** Um coeficiente de correlação de Pearson próximo de 1 ou -1 sugere uma forte relação linear, enquanto valores próximos de 0 indicam uma relação fraca ou inexistente.

A correlação de Pearson é útil para entender a força e a direção de uma relação linear entre variáveis, mas não captura relações não lineares.

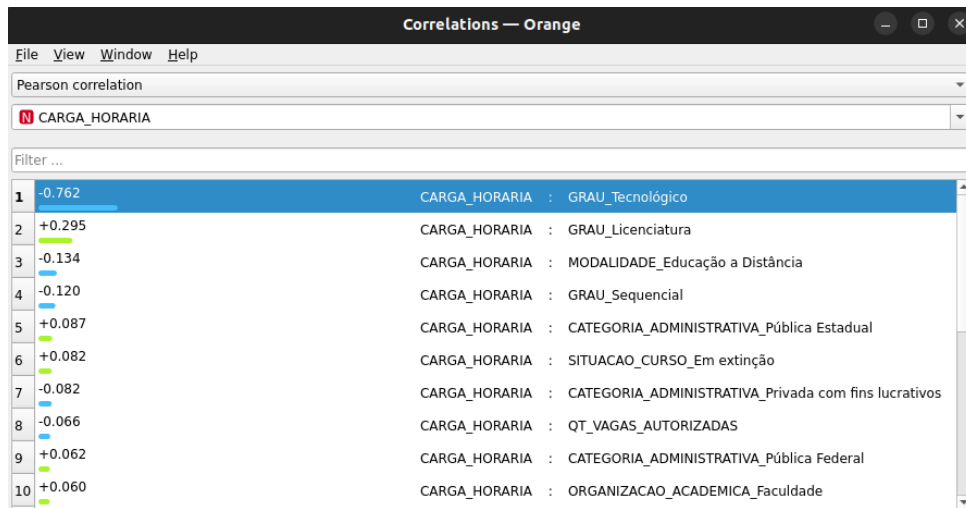


Figura 17 – Correlação de Pearson.

A regressão linear é uma técnica estatística usada para modelar e analisar a relação entre uma variável dependente e uma ou mais variáveis independentes. O objetivo é encontrar a melhor linha reta que se ajusta aos dados, minimizando a soma dos quadrados das diferenças entre os valores reais e os valores previstos pela linha.

Em sua forma mais simples, a regressão linear analisa a relação entre duas variáveis:

- **Variável Dependente (y):** A variável que estamos tentando prever.
- **Variável Independente (x):** A variável usada para prever a variável dependente.

A equação da linha de regressão é:

$$y = \beta_0 + \beta_1 x + \epsilon$$

onde:

- β_0 é o intercepto da linha.
- β_1 é o coeficiente de regressão que representa a inclinação da linha.
- ϵ é o termo de erro.

No gráfico, a regressão linear é representada por uma linha reta que tenta se ajustar da melhor maneira possível aos pontos de dados. O gráfico mostra:

- **Pontos de Dados:** Representam as observações reais da variável dependente e independente.
- **Linha de Regressão:** A linha reta que minimiza a soma dos quadrados das distâncias verticais entre os pontos de dados e a linha.

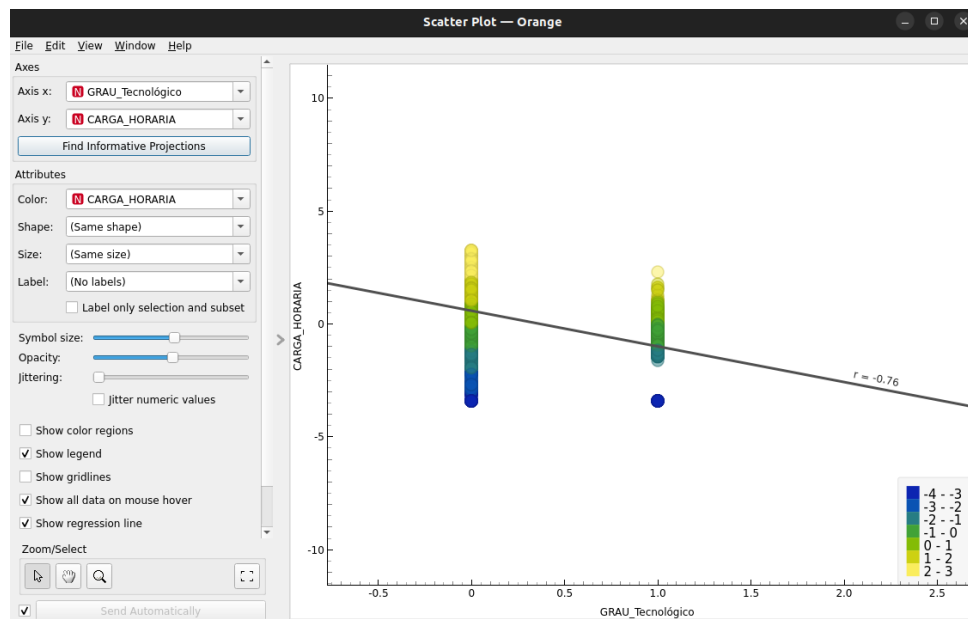


Figura 18 – Gráfico de regressão GRAU por CARGA HORARIA.

4 CONCLUSÃO

Os principais achados do artigo foi um bom desempenho do modelo de classificação, com 79% de acurácia, porém, para o modelo de regressão os erros foram bastante altos. Esses resultados contribuem para lideranças de universidades e faculdades na tomada de decisão em relação a oferta de cursos. Algumas limitações foram encontradas, principalmente em relação ao dataset, onde não encontra-se tantos dados numéricos, dificultando a criação de modelos de predição com uso de regressão.

REFERÊNCIAS

MEC. *Cursos de Graduação do Brasil*. 2022. Disponível em: <https://dadosabertos.mec.gov.br/indicadores-sobre-ensino-superior/item/183-cursos-de-graduacao-do-brasil>.