Fernando Panizzon de Paula

Uso de técnicas de aprendizado de máquina para predição de chuvas na cidade do Rio Grande

Rio Grande, Rio Grande do Sul, Brasil

Fevereiro, 2025

Fernando Panizzon de Paula

Uso de técnicas de aprendizado de máquina para predição de chuvas na cidade do Rio Grande

Trabalho de Conclusão de Curso apresentado ao Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande como requisito parcial para a conclusão do curso de Matemática Aplicada Bacharelado.

Universidade Federal do Rio Grande - FURG

Instituto de Matemática, Estatística e Física - IMEF

Curso de Matemática Aplicada Bacharelado

Orientador: Professora Doutora Raquel da Fontoura Nicolette

Rio Grande, Rio Grande do Sul, Brasil

Fevereiro, 2025

Fernando Panizzon de Paula

Uso de técnicas de aprendizado de máquina para predição de chuvas na cidade do Rio Grande

Trabalho aprovado

Documento assinado digitalmente
gov.br
RAQUEL DA FONTOURA NICOLETTE
Data: 14/02/2025 11:01:51-0300
Verifique em https://validar.iti.gov.br

Profª Doutora Raquel da Fontoura Nicolette
(Orientadora - IMEF- FURG)

Documento assinado digitalmente
gov.br
FELIPE RICARDO SANTOS DE GUSMAO
Data: 10/02/2025 11:24:56-0300
Verifique em https://validar.iti.gov.br

Prof. Doutor Felipe Ricardo Santos de Gusmão
(Avaliador - IMEF - FURG)

Documento assinado digitalmente
gov.br
TIAGO DA CRUZ ASMUS
Data: 10/02/2025 11:50:57-0300
Verifique em https://validar.iti.gov.br

Prof. Doutor Tiago Asmus
(Avaliador - IMEF - FURG)

Rio Grande, 07 de Fervereiro de 2025.

Rio Grande do Sul, Brasil

Agradecimentos

Gostaria de agradecer primeiramente a minha família e amigos por todo o apoio durante essa trajetória.

Quero agradecer principalmente aos meus pais Florinaldo Rodrigues de Paula e Elis Regina Panizzon que possibilitaram a oportunidade de fazer este curso, sem eles eu não seria nada. Quero agradecer às minhas avós Lourdes Panizzon e Idalina Rodrigues de Paula e minha tia Clenilda Strieski, que sempre me incentivaram, mesmo de longe, sempre mandando mensagens pra saber se eu estava bem.

A minha orientadora Raquel da Fontoura Nicolette, por ter aceito passar esse desafio comigo, e agradeço também por todo seu conhecimento e experiência compartilhados. Além disso, agradeço a todos os professores que me deram aula nesses 5 anos cursando Matemática Aplicada.

Aos meus amigos, quero agradecer ao Fernando Rocha Barreto e Paulo Pereira Júnior por sempre estarem comigo, me incentivando do modo de vocês, sei que sempre foi de coração. Aos meus amigos Marcos Max Mello, Amando Matos Ribeiro e Nathalia Maria Schmidtke que nesse tempo que estive fora se fizeram mais presentes, me fazendo ligações e me ajudando com momentos difíceis. Agradeço a toda Sociedade. A todos que sempre me mandaram mensagens ou confraternizaram nas férias.

Ao Selmar Antunes de Oliveira e Sonia Portanova e família que me deram apoio em Rio Grande, e todos que conheci na Fruteira Atlanta.

A todos os meus colegas de curso que sofreram de ansiedade junto comigo antes de uma prova, em especial os que frequentam o laboratório LEA.

E por último, a todas as pessoas que passaram em minha vida e de alguma forma contribuíram nas minhas decisões, que fizeram de alguma forma eu chegar até aqui.

—

“Pode ser difícil agora, mas você deve silenciar esses pensamentos. Pare de contar as coisas que você perdeu, o que se foi se foi. Então, pergunte a si mesmo o que ainda resta para você.”

(Jimbei, One Piece)

Resumo

Este trabalho de conclusão de curso em Matemática Aplicada propos mostrar como os eventos climáticos extremos são perigosos, com foco principal na cidade do Rio Grande - RS, tendo em vista os acontecimentos no ano de 2024 no estado do Rio Grande do Sul, que teve um episódio triste com a alta taxa de chuvas e tempestades durante os meses de abril e maio que levou a recordes de acúmulo de água. Utilizando os métodos de machine learning *Random Forest* e Regressão Logística com auxilio da linguagem R, foi obtido resultados satisfatórios para a predição de precipitação na cidade e os modelos tiveram resultados semelhantes, sendo o melhor a Regressão Logística com 74,4% com a métrica de acurácia e o modelo *Random Forest* com 0,796 com a métrica da curva ROC.

Palavras-chave: Aprendizado de máquina; Chuvas; Eventos extremos; Estátistica ;*Random Forest*; Regressão logística;

Abstract

This final course work in Applied Mathematics proposed to show how extreme weather events are dangerous, with a main focus on the city of Rio Grande - RS, because of the events in the year 2024 in the state of Rio Grande do Sul, which had a sad episode with the high rate of rain and storms during April and May that led to records of water accumulation. Using the machine learning methods *Random Forest* and Logistic Regression with the help of the R language, satisfactory results were obtained for the prediction of precipitation in the city and the models had similar results, the best being Logistic Regression with 74.4% with the accuracy metric and the *Random Forest* model with 0.796 with the ROC curve metric.

Keywords: Machine learning; Rainfall; Extreme events; Statistics; Random Forest; Logistic regression

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Desde os primórdios da humanidade entender como os eventos climáticos funcionam era um desafio, em um primeiro momento tudo era feito de forma mais empírica, observando os sinais da natureza e tentando encontrar algum tipo de padrão. Com o avanço das tecnologias foi possível criar métodos para predição de tais eventos, tendo hoje em dia um amplo aparato científico com a finalidade de evitar grandes desastres, em especial os causados pelas chuvas.

De acordo com uma pesquisa da Organização Meteorológica Mundial (OMM) (UNIDAS, 2021), nos períodos de 1970 e 2019, os desastres naturais correspondem a 50% de todos os desastres registrados, trazendo perdas tanto de vidas quanto econômicas, sendo 45% das mortes registradas no período e 74% de todas as perdas econômicas, causadas por mais de 11 mil eventos extremos com pouco mais de 2 milhões de mortes e 3,47 trilhões de dólares em perdas em todo o mundo, em especial, 91% das mortes foram em países em desenvolvimento. Dos diversos eventos climáticos extremos, os relacionados à chuva como tempestades e enchentes estão no topo da lista dos 10 piores eventos climáticos, causando 577 mil e 58,7 mil mortes respectivamente, ficando atrás apenas das secas que causaram 650 mil mortes, e tendo as tempestades como a maior fonte de perdas econômicas. Mesmo tendo todos esses elevados números, pela existência dos métodos preditivos e de alarmes para aviso da população, o número de mortes foi reduzido a quase um terço entre 1970 e 2019 - caindo de 50 mil nos anos de 1970 para menos de 20 mil nos anos de 2010.

As chuvas têm um papel importante em todo o mundo, elas são responsáveis pela manutenção da terra, ajuda nas lavouras, na geração de energia elétrica, por esses e outros motivos existe uma necessidade de saber de quando irá ocorrer e seu volume para evitar possíveis desastres, como o que ocorreu no estado do Rio Grande do Sul no ano de 2024, onde mais de 450 municípios foram afetados por intensas chuvas e tiveram fortes enchentes, causando a morte de 182 pessoas e deixando um rastro de destruição no custo aproximado de 100 bilhões de reais. As fortes chuvas que ocorreram principalmente no período no final do mês de abril e o mês de maio com acumulados que passam de 1000 mm de chuva (METSUL, 2024), como o caso de Caxias do Sul, na Serra Gaúcha, o acumulado na estação convencional entre 27 de abril e 31 de maio chegou a 1.023,00 mm. No mesmo período outras cidades também tiveram marcas impressionantes, como por exemplo Bento Gonçalves com 961,0 mm, Canela com 967,4 mm e Cambará do Sul com 748 mm e muitas outras cidades no interior gaúcho tiveram marcas significativas, já a capital Porto Alegre no período apenas do mês de maio teve o acumulado de 536,6 mm, sendo o normal 112,9 mm de chuva. Na cidade de Rio Grande foi registrado o total de 316,8 mm de chuva, não

sendo atingida de forma tão assídua como as outras cidades, mas ela se encontra em um ponto estratégico no escoamento da água por estar no ponto de deságue no oceano.

Para diminuir as mortes e danos às cidades são utilizados os dados meteorológicos adquiridos por estações pluviométricas, satélites e radares e aplicados métodos para predição da chuva. Com avanço da tecnologia foram criadas novas técnicas como as de *Machine Learning* (**ML**), em que o computador aprende sobre determinado problema e tenta encontrar respostas de novos problemas, em outras palavras, são utilizadas bases de dados que o algoritmo escolhido aprende a resolver uma situação e após esse treino é apresentada uma nova base de dados para resolução deste mesmo problema.

A matemática desempenha um papel fundamental nos métodos de aprendizado de máquina, fornecendo a base teórica que permite a construção, análise e otimização de algoritmos. A álgebra linear é crucial para o entendimento de operações com vetores e matrizes, que são centrais em muitas técnicas de **ML**, na técnica de Regressão Logística por exemplo as entradas são combinadas linearmente usando pesos, e essa combinação é transformada pela função sigmóide (função que graficamente tem um formato de S) para produzir uma probabilidade, também pode ser usada para calcular as combinações lineares e para a atualização dos pesos durante o treinamento, especialmente em implementações que utilizam métodos como o gradiente descendente. O modelo de *Random Forest* faz a previsão com base na média das previsões (para regressão) ou na votação majoritária (para classificação) das árvores individuais e, nisso a probabilidade desempenha um papel essencial na construção das árvores e na agregação das previsões, como por exemplo a entropia ou o índice de Gini, que são funções probabilísticas, que são utilizadas para decidir as divisões nas árvores.

Dentro deste contexto, o trabalho propõe a utilização de métodos de aprendizado de máquina para entender se eles são eficazes na predição de chuvas, mais especificamente na cidade do Rio Grande. Os métodos utilizados serão o *Random Forest* e Regressão Logística, tanto para averiguar se ambos são boas escolhas para esse objetivo ou não, e entender, se possível, qual é o melhor deles utilizando os dados do INMET (Instituto Nacional de Meteorologia) entre os anos de 2013 e 2024 e descobrir se essa análise ajudaria na prevenção de eventos extremos.

1.1 Revisão bibliográfica

Nesta seção é apresentada uma breve descrição de trabalhos que serão utilizados como base na elaboração deste estudo.

Uso de Regressão Logística para Identificar os Fatores de Risco associados à Ocorrência de Anomalias Congênitas em Recém-nascidos, (SOUZA, 2013). A regressão logística foi aplicada para identificar os fatores de risco associados

à ocorrência de anomalias congênitas em recém-nascidos. O modelo proposto utilizou variáveis como idade materna, escolaridade, uso de corticoides, tipo de parto e os índices APGAR. Os resultados mostraram que o modelo ajustado foi capaz de classificar corretamente mais de 93% dos casos, demonstrando a utilidade do método na predição de malformações congênitas.

Aplicação de regressão logística e modelo de mistura em um estudo sobre clientes inadimplentes de uma empresa de telecomunicações, (HANREJSZKOW A.; STROMBERG, 2013). A regressão logística foi aplicada para modelar a probabilidade de pagamento de clientes inadimplentes de uma empresa de telecomunicações. O modelo identificou os fatores que influenciam a inadimplência, como o perfil do cliente e o tempo até o pagamento. A técnica demonstrou um ajuste satisfatório, permitindo prever o comportamento dos clientes quanto ao pagamento das dívidas.

Previsão de vendas do comércio de varejo com técnicas clássicas e de aprendizado de máquina , (NOSEDA, 2021). Foi utilizada a técnica de *Random Forest* para criar um modelo de aprendizado supervisionado que conseguisse prever o volume de vendas com base em dados históricos e variáveis explicativas relacionadas ao mercado. O modelo mostrou-se robusto para lidar com grandes volumes de dados, identificando padrões complexos que influenciam a variação nas vendas, apresentando um bom desempenho em termos de acurácia.

Análise comparativa entre os algoritmos máxima verssimilança e random forest: classificação dos bosques de manga na área de proteção ambiental de Santa Cruz, Pernanbuco, Brasil, (QUEIROZ, 2021). O algoritmo *Random Forest* foi aplicado para classificar bosques de mangue na Área de Proteção Ambiental de Santa Cruz, Pernambuco. O modelo utilizou uma combinação de imagens de sensoriamento remoto, como dados ópticos do Landsat-5 e índices de vegetação, para melhorar a acurácia da classificação. A técnica *Random Forest* se mostrou eficiente ao apresentar uma acurácia global superior a 95% em todos os conjuntos de dados, demonstrando sua robustez na classificação de diferentes tipos de vegetação.

Algoritmos de machine learning aplicados na ocorrência de chuvas na cidade de Santa Maria, (FACCO M., 2020). Foram utilizadas as técnicas de machine learning Análise Discriminante Linear, Support Vechtor Machines (SVM), Método dos vizinhos mais próximos, Árvores de Classificação e Regressão (CART), Random Forest (RF) para predição de chuvas na cidade de Santa Maria no período de 20 de setembro de 2018 à 20 de setembro de 2019. Os resultados de todos os modelos foram exelentes, sendo o melhor de todos com o modelo Random Forest com 94,84% de acurácia, entretanto o escolhido como melhor foi o CART pelo desempenho computacional.

1.2 Objetivos

Objetivos Gerais

- Implementar um modelo probabilístico para analisar a precipitação na cidade do Rio Grande e realizar a predição das chuvas utilizando os métodos de aprendizado de máquina: Regressão Logística (**RL**) e *Random Forest* (**RF**).

Objetivos Específicos

- Filtrar os dados meteorológicos necessários para o treinamento e teste dos modelos.
- Implementar os algoritmos de **ML** para predição de chuvas.
- Comparar os modelos de previsão e identificar se são adequados e eficientes para a predição de chuvas.
- Validar os resultados obtidos utilizando métricas de desempenho adequadas.

Estrutura do trabalho

Para melhor compreensão do trabalho, ele será dividido como segue:

Capítulo 1 - Introdução: apresenta a motivação, justificativa e objetivos deste trabalho, assim com um breve resumo dos acontecimentos no estado do Rio Grande do Sul no ano de 2024 e o impacto das chuvas.

Capítulo 2 - Fundamentação matemática: discute os conceitos de estatística básica, Machine Learning e métricas de validação dos modelos de **ML**.

Capítulo 3 - Metodologia: neste capítulo será discutida a obtenção dos dados, como serão tratados, como a análise inicial será feita e a ordem de preparação dos modelos de **ML**.

Capítulo 4 - Resultados: neste capítulo serão apresentados os resultados das análises propostas.

Capítulo 5 - Considerações finais: Será feita a comparação dos resultados com outros trabalhos analisados. Um breve resumo do objetivo do trabalho, como possíveis melhorias e conclusão.

2 Fundamentação matemática

A fundamentação matemática e estatística é de grande importâcia na análise de dados, e são indispensáveis para a análise dos resultados dos modelos de **ML** para predição de precipitação na cidade de Rio Grande.

A relação entre Estatística e Machine Learning é fundamental e complementar. A Estatística fornece a base teórica para entender os dados e tomar decisões informadas, enquanto o Machine Learning (**ML**) aplica esses conceitos de maneira prática para resolver problemas complexos. Já a estatística ajuda a explicar os resultados de modelos de **ML**, oferecendo intervalos de confiança e análise de sensibilidade.

2.1 Conceitos básicos de estatística

Estatística é um ramo importante da matemática, que se concentra em métodos para a obtenção de dados, organização, análise e interpretação dos resultados (BUSSAB WILTON DE OLIVEIRA; MORETTIN, 2017).

Na estatística temos o conceito da variável, que é o objeto de estudo da pesquisa realizada, independente da pesquisa. Ela pode ser classificada de duas formas, sendo a quantitativa, quando ela é um valor numérico, como a altura de um indivíduo ou sua idade, e a qualitativa, quando se trata de uma característica, como por exemplo o município de origem ou cor do indivíduo.

Temos os conceitos de população e amostra, a população é o conjunto total de dados utilizado na pesquisa, já a amostra é um subconjunto de dados da população.

Uma das formas de organizar os dados é com o conceito de moda, sendo o valor ou informações que mais se repete na varíavel. Por exemplo, uma turma faz a medição da altura dos alunos e viu que a altura 1,73 m foi a mais recorrente entre os alunos, logo ela é a moda.

2.2 Machine learning

Machine Learning (**ML**) é um subconjunto da inteligência artificial que permite que um sistema aprenda e melhore de maneira autônoma, sem ter sido programado explicitamente para isso, ele tenta identificar um padrão ou reconhecer um objeto.

Os dados da pesquisa são divididos em amostras de treinamento e teste. Os dados de treinamento passam pelo algoritmo para deixar o modelo preciso. Durante o treino,

os dados passam por um processo de ajuste, caso o resultado apresentado seja negativo, o algoritmo é treinado repetidamente até ter uma resposta satisfatória.

Após o modelo treinado, é utilizado a amostra de teste para validar o modelo.

2.3 Tipos de Machine learning

Existem três tipos de **ML**, sendo o supervisionado, não supervisionado e semi-supervisionado (CLOUD, 2024).

O aprendizado supervisionado, o qual tem interferência humana, tem os dados estruturados para encontrar um atributo específico a uma variável resposta.

O aprendizado não supervisionado, sem ajuda humana, não tem dados estruturados e não se sabe qual é a variável resposta e o método deve encontrar uma forma de organizar os dados em grupos, como por exemplo imagens de cachorros e gatos, aprender a fazer o grupo "cachorro" e o grupo "gato" sem exemplos do que são gatos e cachorros, temos como esses métodos de **ML** como exemplo o cluster k-means, clustering hierárquico, entre outros.

Por último, o aprendizado por reforço ou semi-supervisionado, que através de tentativa e erro o modelo aprende através de feedbacks positivos ou negativos se está tendo um resultado satisfatório, por exemplo, um modelo que sabe as regras do jogo de dama e é treinado até o ponto de conseguir ganhar de um humano.

Os métodos de **ML** escolhidos para este trabalho foram o *Random Forest* e Regressão Logística que são do tipo supervisionado.

2.3.1 Random Forest

O *Random Forest* (**RF**) (NOSEDA, 2021), é um modelo de **ML** supervisionado que funciona através de árvores de decisão. Essas árvores começam, por exemplo, com uma pergunta *Vai chover amanhã?*, após isso é feita uma sequência de perguntas relacionadas a esse assunto e essas são chamadas de nós. Observações que atendem aos critérios seguirão o ramo *Sim* e aquelas que não atendem seguirão o caminho alternativo. As árvores de decisão buscam encontrar a melhor divisão para submeter os dados, e geralmente são treinadas através do algoritmo de classificação ou regressão.

Algumas das principais vantagens no uso do **RF** são:

1. Utiliza múltiplas árvores de decisão, reduzindo o risco de *overfitting*, ou seja, quando om modelo se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados, e aumentando a precisão nas previsões;

2. Pode lidar com dados incompletos sem perder muita acurácia;
3. É menos sensível a valores extremos;
4. Funciona bem para problemas de classificação e regressão.

Figura 1 – Exemplo de árvores do método *Random Forest*

2.3.2 Regressão Logística

A Regressão Logística (HANREJSZKOW A.; STROMBERG, 2013) é um modelo de **ML** supervisionado utilizado principalmente para problemas de classificação binária, onde o objetivo é prever a probabilidade de uma observação pertencer a uma das duas classes possíveis. Na Regressão Logística, as variáveis numéricas são usadas como entradas para o modelo linear que é transformado pela função sigmoide para produzir uma probabilidade.

2.4 Modelo matemático de Regressão Logística

Modelos de regressão buscam estabelecer relações entre uma variável resposta e variáveis explicativas. O modelo de Regressão Logística, em particular, se diferencia do modelo de regressão linear quanto à natureza da variável resposta, caracterizada apenas por valores binários ou dicotômicos, usualmente denotados por 1 e 0, com o valor 1 denominado evento de interesse.

O modelo de Regressão Logística (HANREJSZKOW A.; STROMBERG, 2013) fica definido pelo uso da ligação logito, que serve para transformar probabilidades que variam de 0 a 1 em valores na escala de números reais, em um modelo linear generalizado binomial. O modelo de **RL** fica dado por:

$$\text{logit}(P(Y=1)) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{2.1}$$

sendo:

- $P(Y=1)$ é a probabilidade de sucesso (ou da classe 1),
- $\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$ é o logaritmo da razão de chances (logit),
- β_0 é o intercepto,
- $\beta_1, \beta_2, \ldots, \beta_k$ são os coeficientes das variáveis explicativas,
- $X_1, X_2, \ldots, X_k$ são as variáveis explicativas (ou preditoras).

A equação para a probabilidade $P(Y=1)$ é dada por:

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \tag{2.2}$$

2.4.1 Razão de chance

A razão de chance ou possibilidade (HANREJSZKOW A.; STROMBERG, 2013) é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo.

A razão de probabilidade (Odds Ratio) é definida como:

$$\text{OR} = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}}, \tag{2.3}$$

sendo:

- $\text{Odds}_1 = \frac{P_1}{1-P_1}$ é a razão de chances no primeiro grupo,
- $\text{Odds}_2 = \frac{P_2}{1-P_2}$ é a razão de chances no segundo grupo,
- P_1 e P_2 são as probabilidades de sucesso nos grupos 1 e 2, respectivamente.

Para interpretar o resultado temos,

- **Se OR** $= 1$**:** Não há diferença nas chances do evento entre os dois grupos. Isso indica que o evento é igualmente provável em ambos os grupos.

- **Se OR** > 1**:** O evento é mais provável no grupo 1 em comparação com o grupo 2. Por exemplo, uma OR de 2 indica que o evento tem o dobro de chance de ocorrer no grupo 1.
- **Se OR** < 1**:** O evento é menos provável no grupo 1 em comparação com o grupo 2. Por exemplo, uma OR de 0,5 indica que o evento tem o dobro de chance de ocorrer no grupo 2.

2.5 Métricas de análise dos modelos ML

As métricas fazem parte dos elementos estatísticos que envolvem **ML**, com elas é possível validar os mesmos. Neste trabalho que trata de modelos de classificação, temos dois possíveis resultados, acerto ou erro.

Em problemas de classificação binária, predições podem ter quatro possíveis classes (MARIANO, 2021), sendo:

1. Verdadeiro positivo (VP): quando o método diz que a classe é positiva e a classe era realmente positiva;
2. Verdadeiro negativo (VN): quando o método diz que a classe é negativa e a classe era realmente negativa;
3. Falso positivo (FP): quando o método diz que a classe é positiva, mas a classe era negativa;
4. Falso negativo (FN): quando o método diz que a classe é negativa, mas a classe era positiva;

Uma das formas de representação dos resultados de um método de classificação de dados é a matriz de confusão.

Tabela 1 – Matriz de confusão

Possíveis resultados	Vai chover	Não vai chover
Choveu	VP	FP
Não choveu	FN	VN

A matriz de confusão indica a quantidade de ocorrências que o programa teve para cada uma das quatro categorias.

Agora veremos as métricas que podem ser utilizadas para avaliar a qualidade do classificador. São elas, acurácia, curva ROC, brier score, sensibilidade, especificidade.

2.5.1 Acurácia

A acurácia (MARIANO, 2021) é uma métrica simples que mostra o percentual de acerto do modelo. Ela é dada pela razão entre o total de acertos sobre o total de entradas.

$$\text{Acurácia} = \frac{\text{Número de acertos}}{\text{Total de entradas}} \tag{2.4}$$

2.5.2 Curva de ROC

A curva ROC (FAWCET, 2006), do inglês Receiver Operating Characteristic Curve, ou traduzido "Curva Característica de Operação do Receptor" é um gráfico que permite avaliar um classificador binário. Essa visualização leva em consideração a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (especificidade).

Esse gráfico permite comparar diferentes classificadores e definir qual o melhor com base em diferentes pontos de corte. Na prática, quanto mais próximo do topo do eixo Y melhor o classificador.

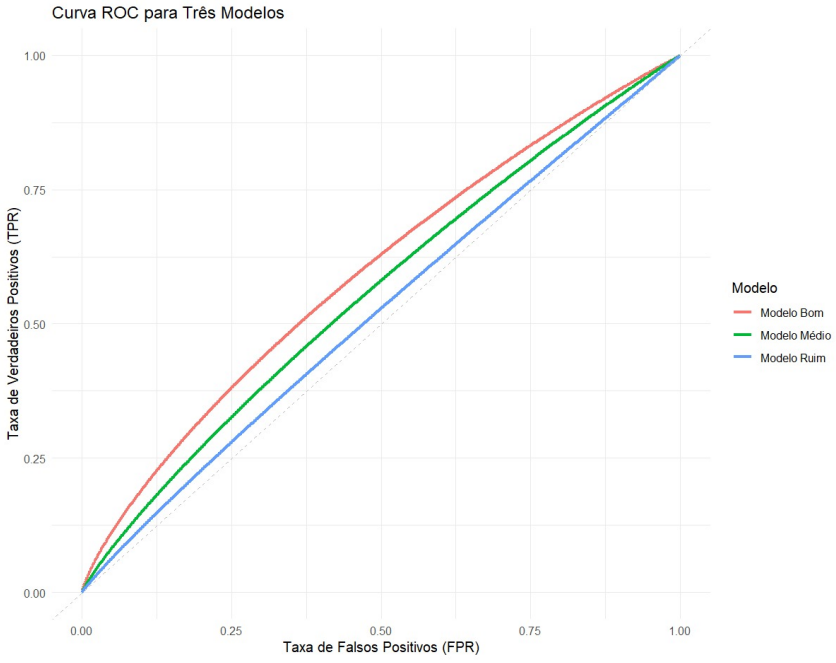

Figura 2 – Exemplo gráfico da curva ROC

Uma curva ROC pode ser avaliada pela métrica AUC (área sob a curva). AUC calcula a área da forma bidimensional formada abaixo da curva que indica a probabilidade de duas previsões serem corretamente ranqueadas. A AUC será um valor entre 0 e 1, quanto mais próximo de 1, melhor a capacidade do modelo em separar as classes.

2.5.3 Brier Score

O Brier Score também é uma métrica estatística utilizada para medir a precisão de previsões probabilísticas. O Brier Score quantifica a diferença entre as probabilidades previstas e os resultados reais, permitindo uma avaliação clara da performance de um modelo preditivo.

Essa métrica varia de 0 a 1, mas diferente das outras quanto mais perto do 0 melhor e quanto mas perto do 1 pior.

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{2.5}$$

onde:

- N **- Número total de previsões:** N representa o número total de observações ou previsões feitas. Ele é usado para calcular a média dos erros quadráticos.
- f_i **- Probabilidade prevista:** f_i é a previsão probabilística para o i-ésimo evento. Ele assume valores no intervalo $[0, 1]$ e indica o grau de confiança na ocorrência do evento.
- o_i **- Valor observado:** o_i é o valor real observado para o i-ésimo evento. No caso de eventos binários, o_i assume valores 0 ou 1, indicando se o evento ocorreu ($o_i = 1$) ou não ocorreu ($o_i = 0$).
- $(f_i - o_i)^2$ **- Erro quadrático:** A diferença $f_i - o_i$ mede o erro da previsão em relação ao valor real. O erro é elevado ao quadrado para garantir que ele seja sempre positivo e para penalizar desvios maiores de forma mais intensa.
- $\sum_{i=1}^{N}(f_i - o_i)^2$ **- Soma dos erros quadráticos:** Representa a soma total dos erros quadráticos de todas as previsões feitas.
- $\frac{1}{N}$ **- Média dos erros quadráticos:** Multiplicamos a soma dos erros quadráticos por $\frac{1}{N}$ para obter a média dos erros quadráticos ao longo de todas as previsões, resultando no valor final do Brier Score.

2.5.4 Sensibilidade

A sensibilidade é a métrica que calcula a taxa de verdadeiros positivos, ou seja, de dias que choveram e o modelo previu corretamente esse resultado.

Essa métrica é calculada pela razão de verdadeiros positivos (VP) sobre a soma de verdadeiros positivos (VP) com os de falsos positivos (FP).

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.6}$$

2.5.5 Especificidade

A especificidade é a métrica que calcula a taxa de verdadeiros negativos, ou seja, de dias que não choveram e o modelo previu corretamente esse resultado.

Essa métrica é calculada pela razão de verdadeiros negativos (VN) sobre a soma de verdadeiros negativos (VN) com os de falsos negativos (FN).

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FN}} \tag{2.7}$$

3 Metodologia

3.1 Base de dados

Os dados utilizados foram retirados do Instituto Nacional de Meteorologia (INMET) (INMET, 2024), tendo uma estação automática localizada na cidade de Rio Grande dentro do campus da Universidade Federal do Rio Grande - FURG, esses dados são disponibilizados diariamente e de hora em hora no site https://bdmep.inmet.gov.br e para o trabalho serão utilizados os dados do ano de 2013 começando no dia 1° de janeiro até 31 de dezembro de 2024.

Figura 3 – Página inicial do INMET

Com os dados em mãos, começa a verificação e organização dos mesmos. Para esse trabalho, os dados adquiridos foram diários. As variáveis que compõem o banco de dados são:

- Data de medição;
- Precipitação total (mm);
- Pressão atmosférica média (mB);
- Temperatura do ponto de orvalho média (°C);
- Temperatura máxima (°C);
- Tempemperatura média (°C);

- Temperatura mínima (°C);
- Umidade relativa do ar, Média (%);
- Vento, rajada máxima (m/s);
- Vento, velocidade média (m/s).

3.2 Suporte Computacional

A linguagem de programação R foi escolhida, pois é excelente para a manipulação dos dados, cálculos estátisticos e aplicação de técnicas de aprendizado de máquina. Aliado a essa linguagem será utilizado o *software* RStudio que é uma ferramente integrada com o R e auxilia para o desenvolvimento do trabalho por causa da sua interface cooperativa com o usuário.

Para este trabalho serão utilizadas as bibliotecas já existentes `tidymodels` e `ranger` para a criação dos modelos de **ML**, treinamento, testes e análise dos resultados, `dyplr` para manipulação dos dados e `ggplot2` para criação dos gráficos.

3.3 Processamento dos dados

A organização do banco de dados começou com a mudança no nome das variáveis para facilitar o manuseio. O segundo passo foi a adição de duas novas colunas: a primeira chamada **Choveu** com os valores *Sim* ou *Nao* para indicar se ocorreu ou não precipitação no dia e a segunda chamada **Chove_amanha**, para indicar se no próximo dia choveu ou não. Exemplificando: supondo que dos dias 01, 02 e 03, choveu apenas no dia 02, a coluna **Choveu** será *Nao*, *Sim*, *Nao*, respectivamente, sendo a coluna **Chove_amanha** será *Sim*, *Nao*, *Nao* (considerando que no dia 04 também não choveu).

3.4 Análise dos dados

Na análise inicial dos dados foi utilizada a forma gráfica de distribuição de densidade nas variáveis quantitativas, o gráfico mostra duas curvas que representam a distribuição da variável quantitativa para a variável resposta.

A sobreposição indica o grau de distinção entre elas, com isso podemos verificar a previsibilidade das mesmas. Quando as curvas têm pouca sobreposição significa que pode ser uma boa variável preditora para a variável resposta; já se as curvas têm muita sobreposição, a variável sozinha não consegue distinguir bem entre as categorias da variável resposta. Outras informações, as quais esses gráficos podem proporcionar, são sobre a altura dos picos (moda). Um alto grau de separação entre os picos indica que a variável

pode ser boa para diferenciar as classes. A largura da curva, sendo mais estreita, significa menor variância dentro de cada classe e curvas largas tem maior variância.

Temos como exemplo o gráfico abaixo, que demonstra as duas curvas em relação as classes da minha variável resposta **Chove_amanha**, sendo elas **Sim** e **Nao**. A partir delas é feita a análise para identificar se é ou não uma boa variável preditora.

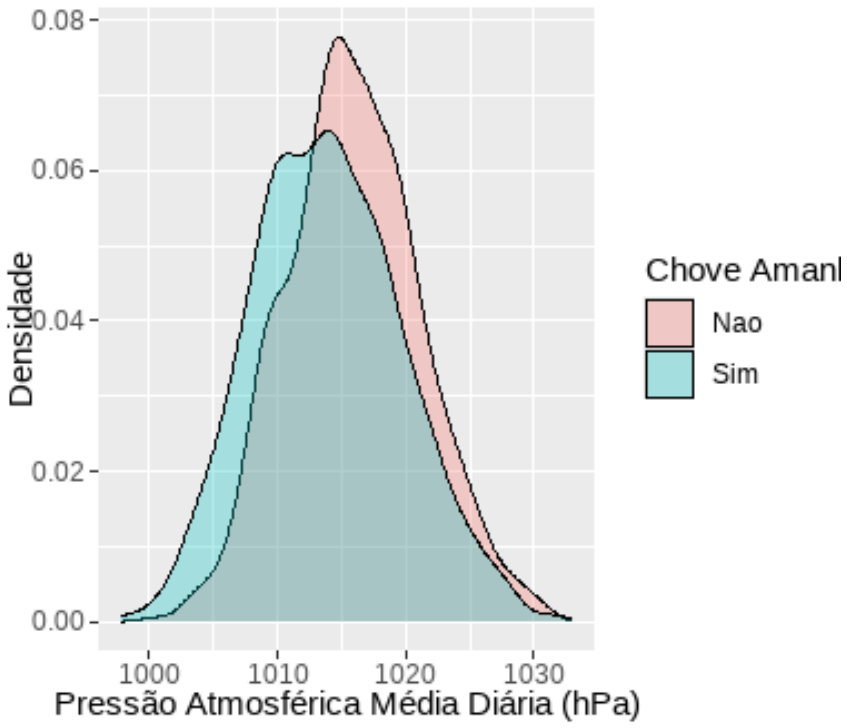

Figura 4 – Exemplo de gráfico de densidade

3.5 Aplicação dos modelos de Machine Learning

Utilizando a biblioteca `tidymodels` do R, faremos duas análises:

- A **Análise 1** será dividindo a população de dados (de 2013 a 2024) em duas amostras: o conjunto de treino (70%) e o conjunto de teste (30%), a partir desse momento essa base teste será chamada de base teste primária.
- Na **Análise 2** os dados de 2013 a 2023 serão utilizados como base treino e os dados de 2024 como base teste, para avaliar uma previsão futura.

A base treino será utilizada para treinar os modelos de **ML**, e a base teste será aplicada após o treino para validar se o modelo é viável ou inviável na previsão de precipitação.

3.5.1 V-folds

Para amenizar um possível viés no resultado final, será utilizada também a separação de dados pela função v-folds, que é utilizada para validação cruzada. Os dados de treino são divididos em partes menores e aproximadamente iguais, para nossa análise teremos $v = 10$ folds.

A validação cruzada acontece da seguinte forma: das 10 folds, uma é utilizada como base teste enquanto $v - 1$ folds são usadas como treino, e isso é feito v vezes. Os resultados posteriores são apresentados pela média aritmética entre os testes de modelo. Vale ressaltar que a base de teste não é a mesma citada anteriormente com 30% dos dados, essa base é usada posteriormente para comparação dos resultados entre ela e as da v-folds.

Essa prática tem vantagens como avaliar a generalização do modelo, como cada dado é usado para teste exatamente uma vez, o modelo é avaliado em várias divisões, reduzindo o risco de overfitting. Outra vantagem é o melhor aproveitamento dos dados, já que todos são usados tanto para treino quanto para teste (em momentos diferentes) e evita o viés de uma única divisão treino-teste, por não depender de uma única divisão dos dados possibilitando uma avaliação mais robusta.

3.5.2 Especificação dos modelos

Com as bases de treino e testes prontas, o próximo passo é criar os modelos de **ML** no R, sendo primeiramente o de regressão logística com o motor `glm` que é uma extensão dos modelos lineares tradicionais o qual permite modelar relações entre variáveis dependentes e independentes, mesmo quando a variável dependente não segue uma distribuição normal.

Para o modelo de **RF** definimos a quantidade mínima de variáveis em cada árvore (*mtry*), o número mínimo de registros para a criação de uma árvore (*min_n*) e a quantidade total de árvores utilizadas. Para a obtenção desse valores foi utilizada a função `tune` a qual ajusta esses valores automaticamente durante a validação cruzada para encontrar o valor que melhora o desempenho do modelo sem ocasionar overfitting e auxilia com o ajuste de hiperparâmetros. O modelo escolhido foi o de classificação e o motor escolhido foi o `ranger` do pacote **ranger**.

3.5.3 Fluxo de trabalho

A primeira etapa do fluxo de trabalho é a junção dos dados pré-processados com os modelos, informando que a variável resposta é **Chove_amanha** e que devem ser utilizadas todas as variáveis da base de dados como preditoras.

3.5.3.1 Treinamento dos modelos ML e resultados

No treinamento dos modelos é onde juntamos toda a preparação anterior para a predição dos modelos **ML** em relação a variável resposta e assim poder coletar as métricas principais. O treinamento acontece de forma independente para todas as separações de dados que serão comparadas na análise. Sendo a separação de 70% dos dados comparada com a função v-folds e posteriormente na segunda análise com a base treino de 2013 a 2023.

No modelo **RL** temos apenas um resultado final para cada separação de dados, ou seja, já retorna de forma direta suas métricas. Já no modelo **RF** temos uma variação de respostas, pois no modelo temos imputado *mtry* e *min_n*, que são valores que variam para melhorar o modelo, para a análise foi definida o valor de `grid` = 25, que permite que o processo experimente 25 combinações desses valores, buscando identificar aquela que otimiza uma métrica principal específica e após isso é escolhido os modelos com as melhores métricas.

4 Resultados

Figura 5 – Gráfico da disposição dos dados diários entre os anos de 2013 a 2018 de se choveu ou não choveu.

Figura 6 – Gráfico da disposição dos dados diários entre os anos de 2019 a 2024 de se choveu ou não choveu.

Com os dados filtrados, foi feita a análise de densidade de variáveis (figuras de 06 a 14). Nesta análise foram escolhidas todas as variáveis, além das variáveis precipitação total diária e a data de medição.

Considerando os gráficos, nas figuras 6, 7, 8, 11, 12, 13 e 14 tivemos os picos separados, o que é uma característica que indica que são boas variáveis para diferenciar as classes. Já nas figuras 9 e 10, têm curvas mais largas, o que demonstra uma maior variância, o que também ajuda a diferenciar as classes.

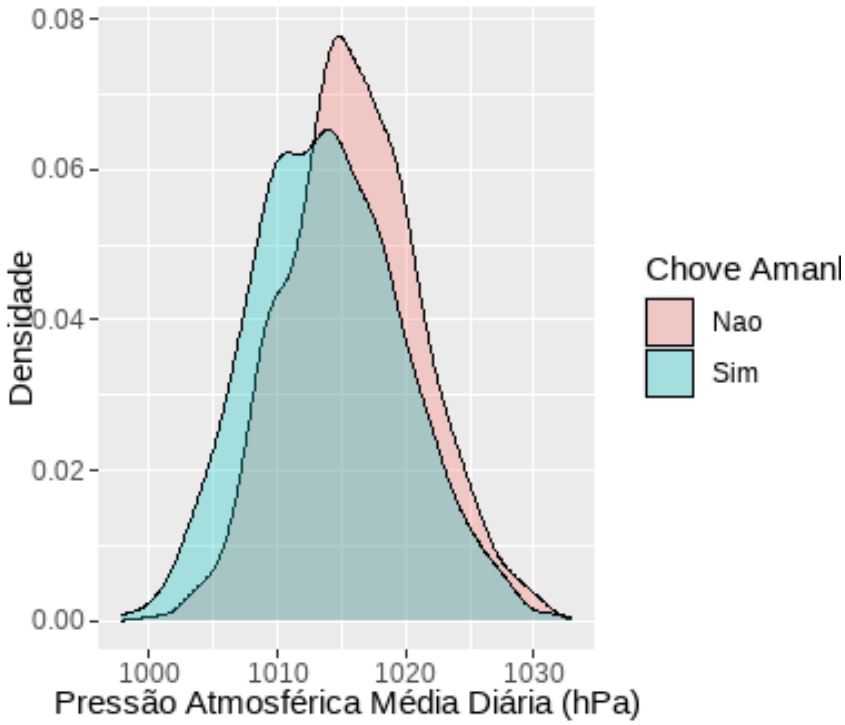

Figura 7 – Gráfico de densidade da correlação entre a pressão atmosférica diária e se choveu ou não choveu.

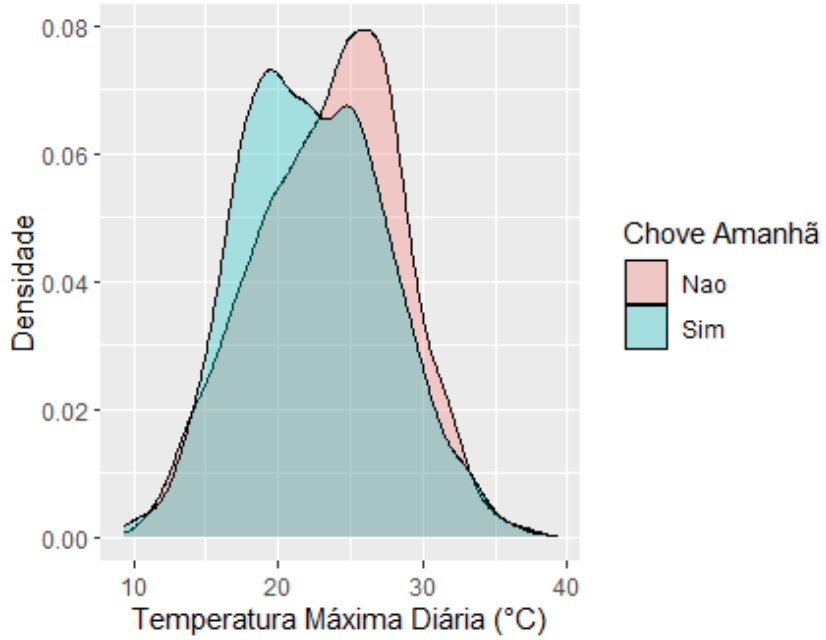

Figura 8 – Gráfico de densidade da correlação entre a temperatura máxima diária e se choveu ou não choveu.

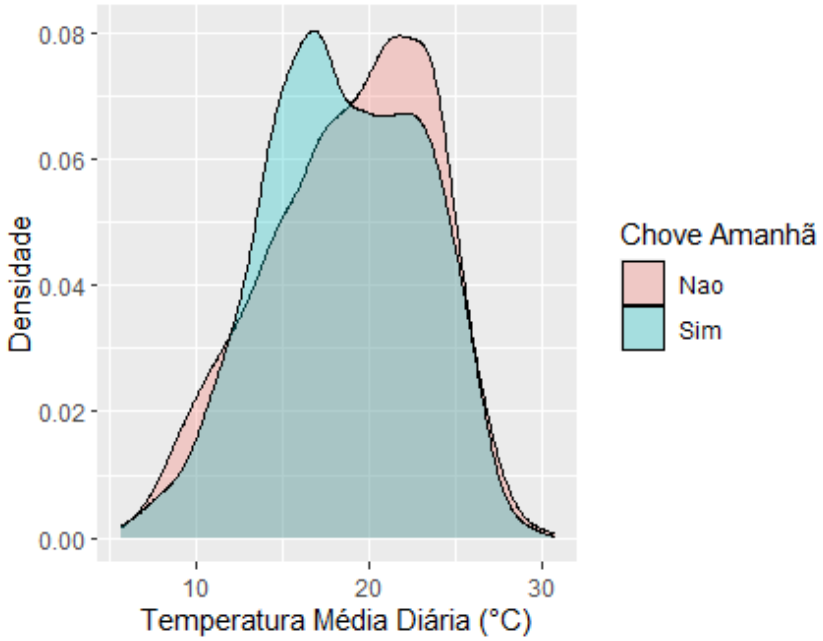

Figura 9 – Gráfico de densidade da correlação entre a temperatura média diária e se choveu ou não choveu.

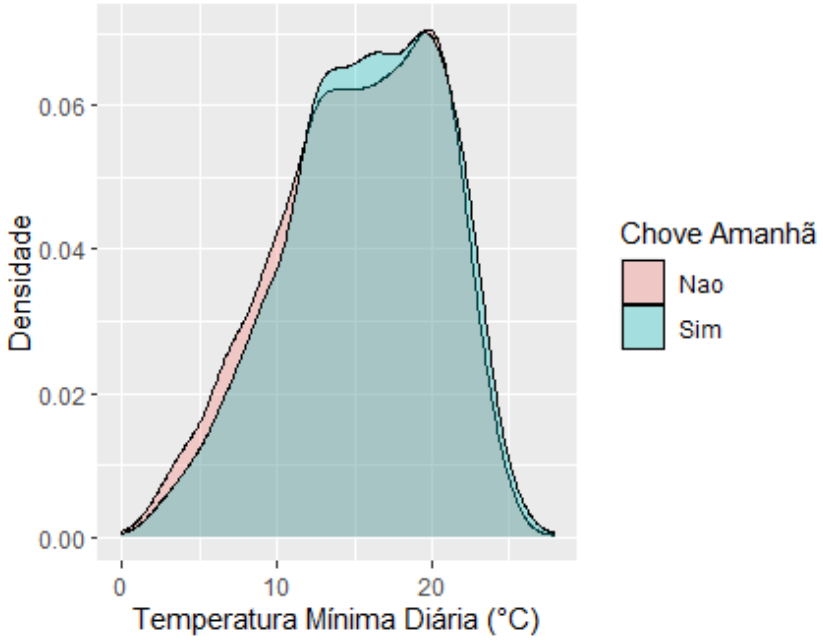

Figura 10 – Gráfico de densidade da correlação entre a temperatura mínima diária e se choveu ou não choveu.

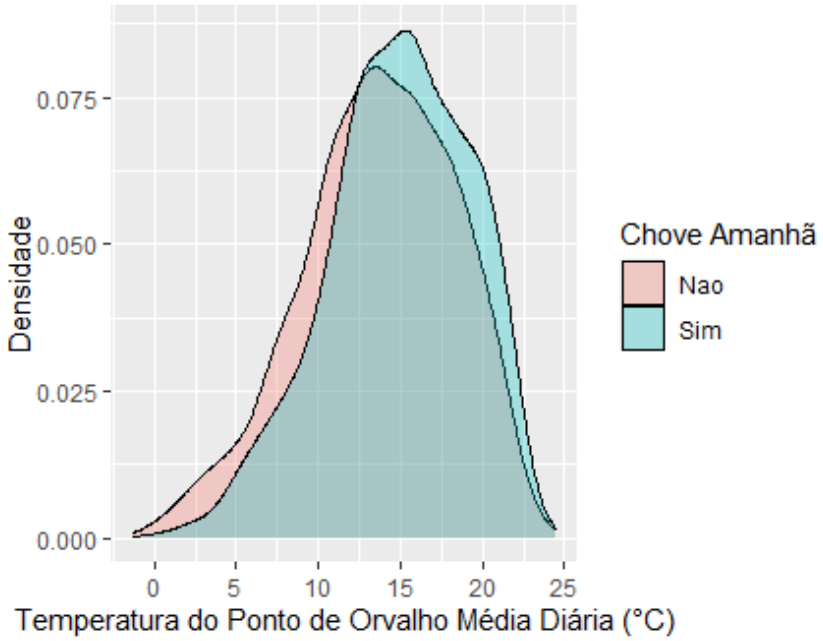

Figura 11 – Gráfico de densidade da correlação entre a temperatura no ponto de orvalho diária e se choveu ou não choveu.

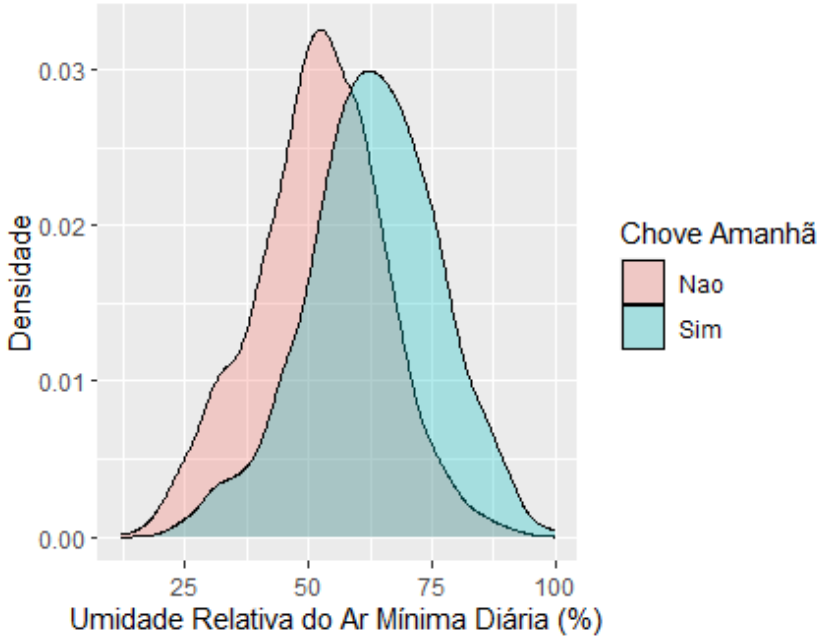

Figura 12 – Gráfico de densidade da correlação entre a umidade mínima diária e se choveu ou não choveu.

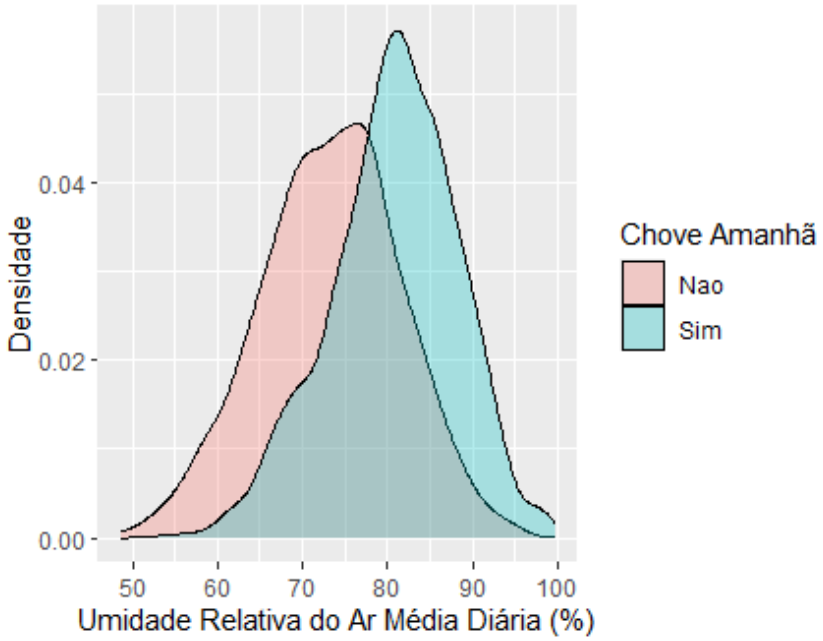

Figura 13 – Gráfico de densidade da correlação entre a umidade média diária e se choveu ou não choveu.

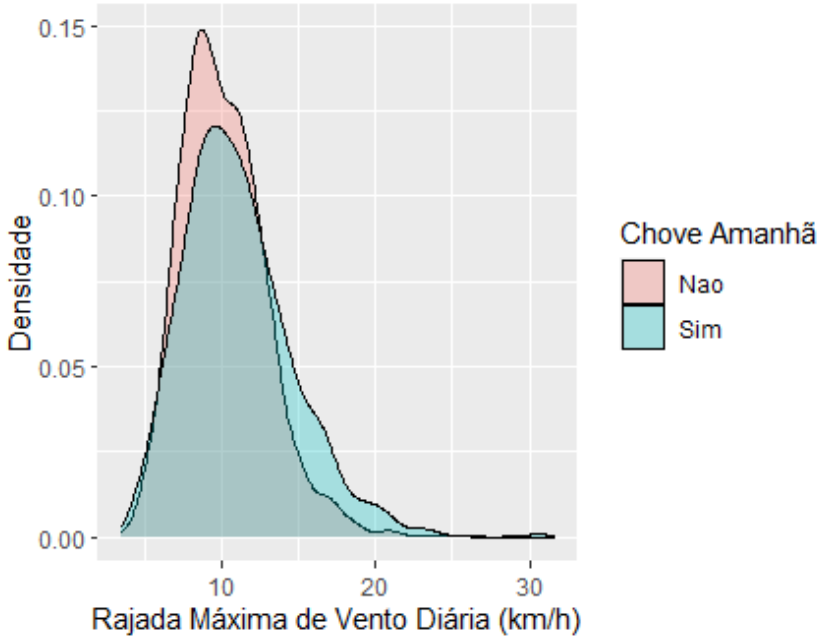

Figura 14 – Gráfico de densidade da correlação entre a rajada máxima diária e se choveu ou não choveu.

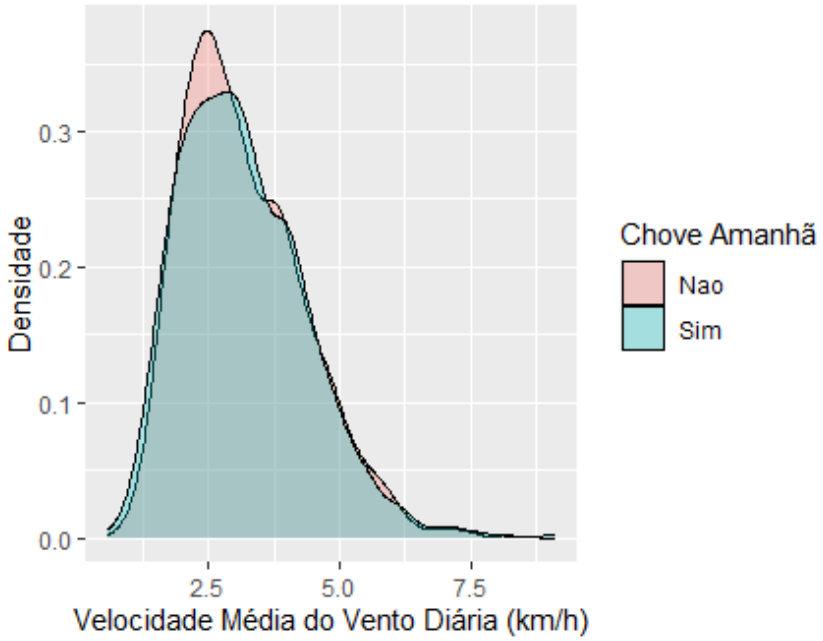

Figura 15 – Gráfico de densidade da correlação entre a velocidade média diária e se choveu ou não choveu.

Após o treino e teste dos modelos para a **Análise 1**, foram obtidos os seguintes resultados dos modelos **RF** e **RF** dispostos nas tabelas 2 à 5.

Tabela 2 – Modelo de Regressão Logística - Análise base teste primária e v-folds

Métrica	Base primária			V-folds		
	Estimador	Média	Número	Estimador	Média	Número
Acurácia	Binário	0,744	1	Binário	0,715	10
Brier Score	Binário	0,183	1	Binário	0,190	10
Curva ROC	Binário	0,794	1	Binário	0,774	10

Tabela 3 – Modelo Random Forest - Análise base teste primária e v-folds

Métrica		Árvores	M-try	M min	Estimador	Média	Número
Acurácia	Base primária	1536	6	37	Binário	0,724	1
Brier Score		1536	6	37	Binário	0,182	1
Curva ROC		1536	6	37	Binário	0,796	1
Acurácia	V-folds	1536	1	14	Binário	0,723	10
Brier Score		1536	1	14	Binário	0,187	10
Curva ROC		1536	1	14	Binário	0,783	10

Os resultados da **Análise 1**, utilizando os dados de 2013 a 2024, mostraram que os modelos **RL** e **RF** são próximos para a previsão de chuva na cidade de Rio Grande.

Comparando os resultados da base teste primária com v-folds, no **RL** tivemos a base primária com 74,4% de acurácia, enquanto v-folds teve 71,5%, com 2,9% de diferença

entre os resultados. Já o modelo **RF** teve a base teste primária com 72,4% de acurácia, enquanto v-folds teve 72,3%, com apenas 0,01% de diferença dos resultados.

Com a matriz de confusão da tabela 4 podemos encontrar as métricas de sensibilidade e especificidade, sendo a sensibilidade de 0,743 para a base primária e 0,708 para a base v-folds. Já a especificidade, temos 0,744 para a base primária e 0,736 para a base v-folds.

Tabela 4 – Matriz de confusão RL - Dados base teste primária e V-folds

	Base primária		V-Folds	
Possíveis resultados	Vai chover	Não vai chover	Vai chover	Não vai chover
Choveu	269	93	77,4	33,1
Não choveu	152	444	48,6	128

Com a matriz de confusão da tabela 5, podemos encontrar as métricas de sensibilidade e especificidade, sendo a sensibilidade de 0,717 para a base primária e 0,717 para a base v-folds. Já a especificidade, temos 0,728 para a base primária e 0,724 para a base v-folds.

Tabela 5 – Matriz de confusão RF - Dados base teste primária e V-folds

	Base primária		V-Folds	
Possíveis resultados	Vai chover	Não vai chover	Vai chover	Não vai chover
Choveu	259	102	76,6	30,2
Não choveu	162	435	49,9	131

Comparando a curva ROC dos resultados da base teste primária com v-folds, no **RL** tivemos a base primária com 0,794, enquanto v-folds teve 0,774, com 0,020 de diferença entre os resultados. Já o modelo **RF** teve a base teste primária com 0,796, enquanto v-folds teve 0,783 de acurácia, com 0,013 de diferença nos resultados.

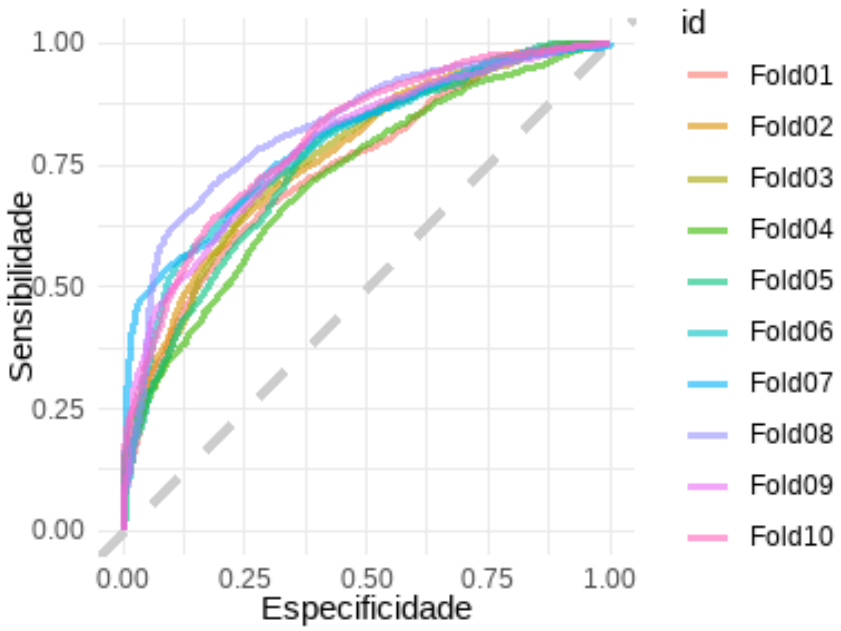

Figura 16 – Gráfico da curva ROC comparando as 10 folds

O modelo **RL** é mais simples do que o **RF**, logo, por serem equivalentes no resultado, podemos considerar que o modelo **RL** é preferível pelo custo computacional. Entretanto, se a base de dados aumentasse de forma considerável e deixasse o problema mais complexo, o **RF** seria a melhor escolha.

Podemos concluir na **Análise 1**, que o modelo **RL** por mais que seja mais simples, foi melhor com a base teste primária com o resultado de 74,4% de acurácia, e o modelo **RF** foi melhor com a curva ROC, obtendo 0,796, entretanto, muito semelhante com o modelo **RL** que marcou 0,794. Para a **Análise 1**, o modelo de **RL** para a previsão de chuva na cidade de Rio Grande foi melhor.

No modelo **RF** as variáveis de entrada foram permutadas de forma aleatória, uma por uma, e na figura abaixo é mostrado o impacto dessa permutação no desempenho do modelo. As variáveis que mais influenciaram foram:

A variável preditora de **umidade relativa do ar** foi a mais impactante, seguida da **temperatura máxima**, **umidade relativa do ar mínima** e **temperatura média**.

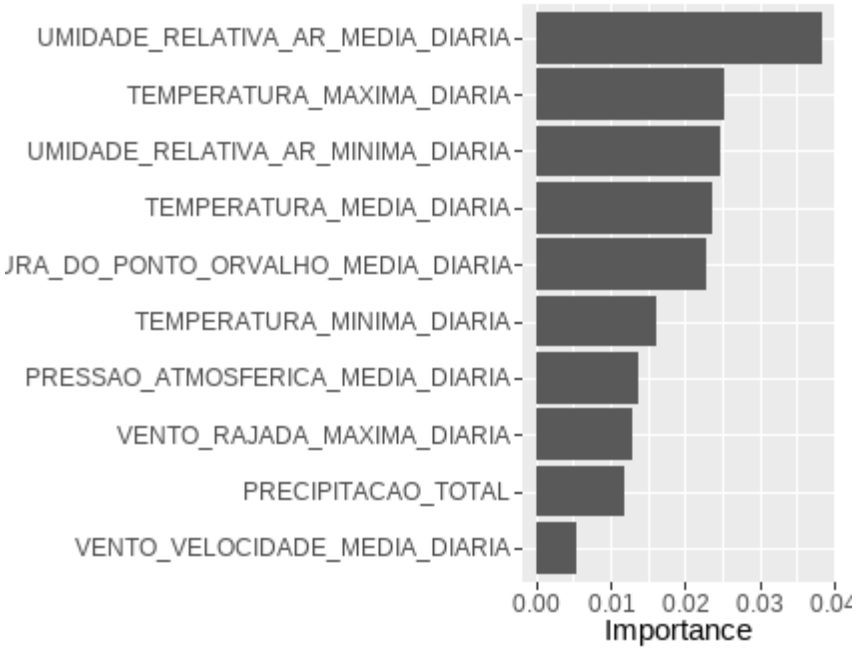

Figura 17 – Gráfico em ordem das variáveis que mais tiveram importância nos modelos de ML

A razão de probabilidade no modelo **RL** foi obtida através da função `tidy()`, com o argumento `(exponentiate = TRUE)` que transforma os valores 0 e 1 em valores numéricos reais, filtrado apenas com p-valor menor que 0.05 para o melhor resultado foi:

Tabela 6 – Modelo de Regressão logística - Análise da razão de probabilidade

Variáveis	Estimativa	P-valor
Vento, rajada máxima	0,832	0,000
Pressão atmosférica	1,066	0,000
Vento, velocidade média	1,388	0,0001
Umidade relativa do ar, média	0,978	0,003

Estes resultados demonstram que:

- Vento, rajada máxima (m/s): A cada unidade a menos, as rajadas máximas de vento contribuem em 16,8% para não chover.
- Pressão atmosférica média (mB): A cada unidade a mais, a pressão atmosférica contribui em 6,6% para chover.
- Vento, velocidade média (m/s): A cada unidade a mais, a velocidade média do vento contribui em 38,8% para chover.
- Umidade relativa do ar, Média (%): A cada unidade a menos, a média da umidade relativa do ar contribui em 2,2% para não chover.

Agora, para a **Análise 2**, após o treino e teste dos modelos, foram obtidos os seguintes resultados dos modelos **RL** e **RF** dispostos nas tabelas 7 a 10.

Tabela 7 – Modelo de Regressão Logística - Análise de predição do ano 2024 com base teste primária e V-folds

Métrica	Base primária			V-folds		
	Estimador	Média	Número	Estimador	Média	Número
Acurácia	Binário	0,744	1	Binário	0,726	10
Brier Score	Binário	0,183	1	Binário	0,185	10
Curva ROC	Binário	0,794	1	Binário	0,782	10

Tabela 8 – Modelo Random Forest - Análise predição do ano 2024 com base teste primária e V-folds

Métricas		Árvores	M-try	M min	Estimador	Média	Número
Acurácia	Base primária	1687	6	37	Binário	0,727	1
Brier Score		1687	6	37	Binário	0,182	1
Curva ROC		1687	6	37	Binário	0,795	1
Acurácia	V-folds	1536	6	37	Binário	0,733	10
Brier Score		1536	6	37	Binário	0,183	10
Curva ROC		1536	6	37	Binário	0,790	10

Para a **Análise 2**, utilizando os dados de 2013 a 2023 como base de treino e 2024 como base de teste para predição de chuva na cidade de Rio Grande, mostraram que os modelos **RL** e **RF** são próximos para a previsão de chuva na cidade de Rio Grande.

Comparando o resultado da base teste primária com v-folds, no **RL** tivemos que a base primária foi melhor com 74,4% de acurácia, enquanto v-folds teve 72,6%, com 1,8% de diferença entre os resultados. Já o modelo **RF** teve a base teste primária com 72,7% de acurácia, enquanto v-folds teve 73,3%, com apenas 0,6% de diferença dos resultados.

Com a matriz de confusão da tabela 9, podemos encontrar as métricas de sensibilidade e especificidade, sendo a sensibilidade de 0.743 para a base primária e 0.708 para a basee v-folds. Já a especificidade, temos 0.745 para a base primária e 0.735 para a base v-folds.

Tabela 9 – Matriz de confusão RL - Predição com os dados base teste primária e V-folds

	Base primária		V-Folds	
Possíveis resultados	Vai chover	Não vai chover	Vai chover	Não vai chover
Choveu	269	93	91,4	37,6
Não choveu	152	444	58,5	163

Com a matriz de confusão da tabela 10, podemos encontrar as métricas de sensibilidade e especificidade, sendo a sensibilidade de 0,713 para base primária e 0,721 para base v-folds. Já a especificidade, temos 0,735 para a base primária e 0,740 para base v-folds.

Tabela 10 – Matriz de confusão RF - Predição com os dados base teste primária e V-folds

	Base primária		V-Folds	
Possíveis resultados	Vai chover	Não vai chover	Vai chover	Não vai chover
Choveu	266	107	92	35,6
Não choveu	155	430	57,9	165

Comparando a curva ROC dos resultados da base teste primária com v-folds, no **RL** tivemos a base primária com 0,794, enquanto v-folds teve 0,782 , com 0,012 de diferença entre os resultados. Já o modelo **RF** teve a base teste primária com 0,795, enquanto v-folds teve 0,790 de acurácia, com apenas 0,005 de diferença nos resultados.

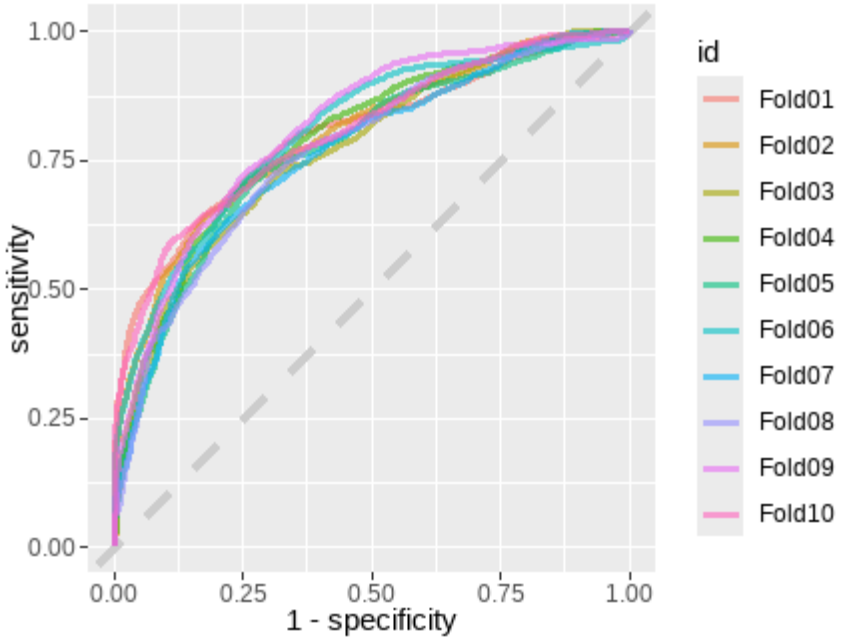

Figura 18 – Gráfico da curva ROC comparando as 10 folds

O resultado do modelo **RL** para a acurácia foi igual na **Análise 1** e **Análise 2**, logo a razão de probabilidade foi igual também.Novamente,e o modelo **RL** com a base de teste primária foi a melhor,r com 74,4% para a acurácia e **RF** para curva ROC com 0,795.

No modelo **RF**, as variáveis que mais influenciaram foram:

As variáveis preditoras de **umidade** foram as mais impactantes nos modelos, com a umidade relativa do ar média sendo mais presente do que na **Análise 1**. Já as de **temperatua** foram menos importantes para a predição.

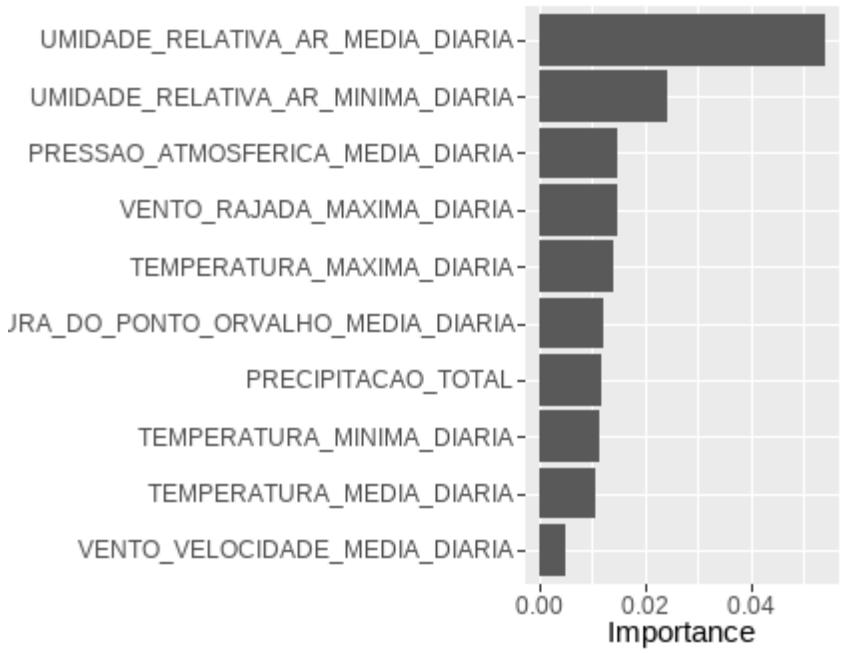

Figura 19 – Gráfico em ordem das variáveis que mais tiveram importância nos modelos de ML

5 Considerações finais

O presente estudo teve como objetivo a implementação dos modelos de Machine Learning, *Random Forest* e Regressão Logística, para a predição de chuvas na cidade de Rio Grande.

Com base nos dados obtidos, tivemos como o melhor resultado o de 74,4% com o modelo de **RL** para a métrica de acurácia e o modelo **RF** 0.796 para a métrica da curva ROC. Além disso, para a cidade do Rio Grande, a variável que mais influenciou foi a umidade relativa do ar média, o que se mostra coerente dada a localização dela no Brasil.

Comparando com os resultados de (FACCO M., 2020), que também utilizou uma cidade do sul do Brasil, Santa Maria - RS, e analisou o período de 20 de setembro de 2018 a 20 de setembro de 2019, mas com o diferencial nos dados, usando de hora em hora. No trabalho o modelo **RF** obteve a acurácia de 94,84%, que se mostra muito superior ao nosso resultado do **RF** com 72,4%. Comparando os melhores resultados, no trabalho sobre Santa Maria foi escolhido o modelo **CART** com a acurácia de 94,74%, pelo seu desempenho computacional, destronando o **RF**. Já neste trabalho, o melhor resultado foi com o modelo **RL** com 74,4%. Isso também demonstra que é possível o melhoramento dos modelos apresentados para que se tenha resultados melhores.

O modelo de **RL** se mostrou por pouco superior neste trabalho, entretanto, o modelo não deve ser considerado absoluto na predição de precipitação diária. Para cada região existe uma base de dados, que mostra condições climáticas distintas da cidade do Rio Grande. Sendo assim, a utilização de outros modelos de **ML**, ou o aumento da área, considerando as cidades de São José do Norte e Pelotas, podem melhorar as métricas dos modelos testados. Outras formas de melhorar os modelos seriam a utilização de técnicas para imputação de dados, aumento da base de dados, consideração de eventos extremos que interferem na cidade, como o El Niño e La Niña.

De forma geral, este trabalho vem para mostrar uma alternativa na predição de chuvas. Com um modelo bem treinado e testado, pode-se tornar uma ferramenta poderosa para a cidade, tendo em vista os acontecimentos da enchente de 2024.

Portanto, podemos concluir que os métodos são válidos e pertinentes para a predição de chuva na cidade do Rio Grande, e abrem caminhos para a forma de análise deste problema. Assim, este trabalho contribui para o uso de **ML** é válido para previsão de chuvas.

Referências

BUSSAB WILTON DE OLIVEIRA; MORETTIN, P. A. *Estatística Básica*. [S.l.]: "Saraiva Educação", 2017. Citado na página 13.

CLOUD, G. *Google Cloud*. 2024. "https://cloud.google.com/learn/what-is-machine-learning?hl=pt-BR.". Citado na página 14.

FACCO M., C. M. A. d. A. d. V. D. d. S. S. R. B. . B. C. Algoritmos de machine learning aplicados na ocorrência de chuvas na cidade de santa maria. *"Ciência E Natura, 42, e28."*, 2020. Citado 2 vezes nas páginas 11 e 40.

FAWCET, T. *An introduction to roc analysis*. [S.l.]: "Pattern Recogn. Lett", 2006. Citado na página 18.

HANREJSZKOW A.; STROMBERG, E. Aplicação de regressão logística e modelo de mistura em um estudo sobre clientes inadimplentes de uma empresa de telecomunicações. *Trabalho de Conclusão de Curso (Estatística) – Universidade Federal do Paraná, Curitiba*, 2013. Citado 3 vezes nas páginas 11, 15 e 16.

INMET, P. *INSTITUTO NACIONAL DE METEOROLOGIA (INMET)*. 2024. "https://portal.inmet.gov.br. Acesso em: 10 set. 2024.". Citado na página 21.

MARIANO, D. Métricas de avaliação em machine learning. *BIOINFO – Revista Brasileira de Bioinformática. Edição 01*, 2021. Citado 2 vezes nas páginas 17 e 18.

METSUL. *Chuva que levou às enchentes no Rio Grande do Sul superou 1000 mm*. 2024. "https://metsul.com/chuva-que-levou-as-enchentes-no-rio-grande-do-sul-superou-1000-mm/. Acesso em: 10 set. 2024.". Citado na página 9.

NOSEDA, F. D. D. Previsão de vendas utilizando aprendizado de máquina. *Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Computação) – Universidade Tecnológica Federal do Paraná, Curitiba*, 2021. Citado 2 vezes nas páginas 11 e 14.

QUEIROZ, V. D. B. e. Análise comparativa entre os algoritmos máxima verossimilhança e random forest: classificação dos bosques de mangue da Área de proteção ambiental de santa cruz, pernambuco, brasil. *Trabalho de Conclusão de Curso (Engenharia Cartográfica e de Agrimensura) – Universidade Federal de Pernambuco, Recife*, 2021. Citado na página 11.

SOUZA, L. D. A. d. Uso de regressão logística para identificar os fatores de risco associados à ocorrência de anomalias congênitas em recém-nascidos. *37 p. Monografia (Bacharelado em Estatística) – Universidade Federal da Paraíba, João Pessoa*, 2013. Citado na página 10.

UNIDAS, N. *Desastres naturais foram responsáveis por 45% de todas as mortes nos últimos 50 anos, mostra OMM*. 2021. <https://brasil.un.org/pt-br/142679-desastres-naturais-foram-responsÃąveis-por-45-de-todas-mortes-nos-Ãžltimos-50-anos-most
Acesso em: 10 set. 2024. Citado na página 9.

A Anexo - Código das análises

```
#Criação do novo BD
dados_df<- dados %>% select(Chove_Amanha,
                            Data_Medicao,
                            PRECIPITACAO_TOTAL,
                            PRESSAO_ATMOSFERICA_MEDIA_DIARIA,
                            TEMPERATURA_DO_PONTO_ORVALHO_MEDIA_DIARIA,
                            TEMPERATURA_MEDIA_DIARIA,
                            TEMPERATURA_MAXIMA_DIARIA,
                            TEMPERATURA_MINIMA_DIARIA,
                            UMIDADE_RELATIVA_AR_MEDIA_DIARIA,
                            UMIDADE_RELATIVA_AR_MINIMA_DIARIA,
                            VENTO_RAJADA_MAXIMA_DIARIA,
                            VENTO_VELOCIDADE_MEDIA_DIARIA)%>%

  mutate_if(is.character, as.factor) %>%
  drop_na() %>%
  mutate(Chove_Amanha = relevel(Chove_Amanha, ref = "Sim"))
glimpse(dados_df)
levels(dados_df$Chove_Amanha)
```

```
#Separação em base treino e teste
set.seed(1232)
dados_split <- initial_split(dados_df, strata = Chove_Amanha)
dados_train <- training(dados_split)
dados_test <- testing(dados_split)
dados_split

#Criação das v-folds
dados_cv <- vfold_cv(dados_train, strata = Chove_Amanha)
dados_cv

#Criação dos modelos RL e RF
### Logistic Regression
dadosglm_spec <- logistic_reg() %>%
  set_engine("glm")

dadosglm_spec
```

```
### Random Forest
dadosrf_spec <- rand_forest(
  mtry = tune(),
  trees = tune(),
  min_n = tune()
) %>%
  set_mode("classification") %>%
  set_engine("ranger")

dadosrf_spec

#Criação do Workflow

dados_wf <- workflow() %>%
  add_formula(Chove_Amanha ~ .)

dados_wf

### Parallel Processing

doParallel::registerDoParallel()

#Aplicação do modelo RL para v-folds e final
dadosglm_rs <- dados_wf %>%
  add_model(dadosglm_spec) %>%
```

```r
#Aplicação do modelo RL para v-folds e final
dadosglm_rs <- dados_wf %>%
  add_model(dadosglm_spec) %>%
  fit_resamples(
    resamples = dados_cv,
    control = control_resamples(save_pred = TRUE)
  )

dadosglm_rs

collect_metrics(dadosglm_rs)

dadosglm_rs %>%
  collect_predictions() %>%
  sensitivity(Chove_Amanha, .pred_class)

dadosglm_rs %>%
  collect_predictions() %>%
  specificity(Chove_Amanha, .pred_class)

dadosglm_rs %>%
  conf_mat_resampled()
```

```r
dados_final <- dados_wf %>%
  add_model(dadosglm_spec) %>%
  last_fit(dados_split)

dados_final #comparação do teste 30%

dados_final %>% collect_metrics()

## Confusion matrix on the Test data

dados_final %>%
  collect_predictions() %>%
  conf_mat(Chove_Amanha, .pred_class)

dados_tune_wf <- dados_wf %>%
  add_model(dadosrf_spec)

dados_tune_wf

#aplicação do modelo RF para v-folds e final
dadosrf_rs_tune <- tune_grid(
  object = dados_tune_wf,
  resamples = dados_cv,
  grid = 25,
  control = control_resamples(save_pred = TRUE)
)
```

```r
### Results of Tuning
dadosrf_rs_tune

### Check Accuracy and AUC after tuning
dadosrf_rs_tune %>%
  collect_metrics()

autoplot(dadosrf_rs_tune)

best_acu <- select_best(dadosrf_rs_tune, metric = "accuracy")
best_acu

dadosrf_rs_tune %>%
  collect_metrics() %>%
  filter(mtry == best_acu$mtry, min_n == best_acu$min_n)

### Also check which model has the highest AUC

best_auc <- select_best(dadosrf_rs_tune, metric = "roc_auc")
best_auc

dadosrf_rs_tune %>%
  collect_metrics() %>%
  filter(mtry == best_auc$mtry, min_n == best_auc$min_n)
```

```r
dadosrf_final <- finalize_model(
  dadosrf_spec,
  best_acu
)

dadosrf_rs_tune %>%
  collect_predictions() %>%
  sensitivity(Chove_Amanha, .pred_class)

dadosrf_rs_tune %>%
  collect_predictions() %>%
  specificity(Chove_Amanha, .pred_class)

conf_matrix_resampled <- dadosrf_rs_tune %>%
  conf_mat_resampled(parameters = best_acu)

# Visualizar a matriz de confusão
conf_matrix_resampled


dadosrf_final #### This is the Best Model from
```

```r
dadosrf_final #### This is the Best Model from

final_res <- dados_wf %>%
  add_model(spec = dadosrf_final) %>%
  last_fit(dados_split)

final_res %>%
  collect_metrics()

final_res %>%
  collect_predictions() %>%
  conf_mat(Chove_Amanha, .pred_class)

# Extraindo o modelo final
final_model <- extract_workflow(final_res) %>%
  extract_fit_parsnip()

# Verificar as especificações do modelo
final_model$fit
```

Universidade Federal do Rio Grande – FURG
Instituto de Matemática, Estatística e Física
Curso de Bacharelado em Matemática Aplicada
Av. Itália km 8 Bairro Carreiros
Rio Grande-RS CEP: 96.203-900 Fone (53)3293.5411
e-mail: imef@furg.br Sítio: www.imef.furg.br

Ata de Defesa de Monografia

No sétimo dia do mês de fevereiro de 2025, às 14h, no Auditório do IMEF, foi realizada a defesa do Trabalho de Conclusão de Curso do acadêmico **Fernando Panizzon de Paula**, intitulada **"Uso de técnicas de aprendizado de máquina para predição de chuvas na cidade do Rio Grande"**, sob a orientação da Profa. Dra. Raquel da Fontoura Nicolette, deste instituto. A banca avaliadora foi composta pela Prof. Dr. Tiago Asmus (IMEF/FURG) e pelo Prof. Dr. Felipe Ricardo Santos de Gusmão (IMEF/FURG). O candidato foi aprovado por unanimidade. Na forma regulamentar, foi lavrada a presente ata, que é abaixo assinada pelos membros da banca, na ordem acima relacionada.

gov.br Documento assinado digitalmente
RAQUEL DA FONTOURA NICOLETTE
Data: 14/02/2025 11:01:51-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Raquel da Fontoura Nicolette
Orientadora

gov.br Documento assinado digitalmente
FELIPE RICARDO SANTOS DE GUSMAO
Data: 10/02/2025 11:23:06-0300
Verifique em https://validar.iti.gov.br

Profº Dr.Felipe Ricardo Santos de Gusmão

gov.br Documento assinado digitalmente
TIAGO DA CRUZ ASMUS
Data: 10/02/2025 11:50:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Tiago Asmus