UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANDREI POCHMANN KOENICH

BRUNO FERREIRA AIRES

FELIPE KAISER SCHNITZLER

FELIPE SOUZA DIDIO

TURMA U

RELATÓRIO – TRABALHO II

Disciplina: Aprendizado de Máquina

Professor: Anderson Rocha Tavares

Porto Alegre, agosto de 2024.

1 COLETA DE DADOS E IDENTIFICAÇÃO DO PROBLEMA

Os dados utilizados na elaboração do trabalho prático estão disponíveis em https://www.kaggle.com/datasets/nikhil7280/weather-type-classification. Esse conjunto de dados foi elaborado com o propósito de incluir características relacionadas ao clima como atributos, a fim de determinar se essas características fazem com que o dia possa ser classificado como "nublado", "nevoso", "ensolarado" ou "chuvoso" (*cloudy*, *snowy*, *sunny* ou *rainy*, respectivamente). No total, existem 13200 instâncias diferentes, com a presença de *outliers*.

2 ANÁLISE EXPLORATÓRIA DOS DADOS

Abaixo, seguem as descrições de cada um dos atributos considerados na elaboração dos modelos, junto à sua importância na tarefa da previsão do tempo.

Os atributos numéricos são:

- **Temperatura (*Temperature*):** indicada em graus Celsius, oscilando entre valores de frio extremo e calor extremo, a temperatura afeta diretamente o tipo de precipitação e a sensação térmica. Extremas temperaturas influenciam a formação de nuvens e tempestades.
- **Umidade (*Humidity*):** indicada em valor percentual, a umidade indica a quantidade de vapor de água no ar. Altos níveis de umidade estão associados a maior probabilidade de chuva, neblina e sensação de abafamento.
- **Velocidade do Vento (*Wind Speed*):** indicada em quilômetros por hora, a velocidade do vento influencia a sensação térmica e pode indicar condições de tempestades. Ventos fortes podem dispersar nuvens e influenciar a distribuição de sistemas meteorológicos.
- **Precipitação Percentual (*Precipitation Percentage*):** indicada em valor percentual. A precipitação percentual indica a probabilidade de ocorrência de chuva ou neve. Valores altos sugerem maior chance de chover ou nevar.
- **Pressão Atmosférica (*Atmospheric Pressure*):** medida em hPa (hectopascal). A pressão atmosférica influencia o movimento das massas de ar. Pressões baixas estão associadas a tempestades e chuvas, enquanto pressões altas geralmente indicam tempo claro e seco.
- **Índice Ultravioleta (*UV Index*):** utilizado para medir a intensidade da radiação ultravioleta do sol. Altos índices ultravioleta são comuns em dias ensolarados e indicam maior risco de exposição solar.
- **Visibilidade (*Visibility*):** medida em quilômetros. A visibilidade reflete a clareza do ar e pode indicar condições de neblina, fumaça ou tempestades de areia. Baixa visibilidade é comum em dias chuvosos ou nevosos.

Os atributos categóricos são:

- **Nebulosidade (*Cloud Cover*):** a cobertura de nuvens afeta a quantidade de luz solar recebida e pode indicar condições de tempo nublado ou ensolarado. É crucial para prever a insolação e a ocorrência de precipitação. Os valores possíveis no conjunto de dados são “nublado”, “parcialmente nublado” e “limpo” (*overcast, partly cloudy* e *clear,* respectivamente).
- **Estação (*Season*):** a estação do ano afeta padrões climáticos sazonais, como temperatura e precipitação, sendo fundamental para a previsão de eventos climáticos típicos de cada estação. Os valores possíveis no conjunto de dados são “primavera”, “verão”, “outono” e “inverno” *(spring, summer, autumn* e *winter,* respectivamente).
- **Local (*Location*):** o tipo de local afeta as condições climáticas locais devido a fatores geográficos específicos que influenciam temperatura, vento e precipitação. Os valores possíveis no conjunto de dados são “montanhoso”, “no litoral” e “no interior” *(mountain, inland* e *coastal, respectivamente*).
- **Tipo do Clima (*Weather Type*):** atributo alvo da classificação, indicando o tipo de tempo. Ajuda a sintetizar todas as outras variáveis para prever o estado do tempo.

Para fazer uso dos atributos categóricos presentes no *dataset*, foram utilizadas as técnicas de pré-processamento indicadas a seguir. As três técnicas foram aplicadas previamente à utilização dos algoritmos que foram utilizados em cada um dos modelos, (descritos na Seção 3), na ordem apresentada abaixo.

- ***Interquartile Range* (IQR):** inicialmente, utiliza-se a técnica IQR, que consiste em uma abordagem estatística usada para identificar e remover *outliers* de um conjunto de dados. Essa técnica fornece uma medida da dispersão central dos dados, de modo a ignorar os valores extremos.

- ***One-hot-encoding*:** a seguir, utiliza-se a técnica de de *one-hot-encoding*, método usado para converter atributos categóricos em atributos numéricos e discretos, de tal forma que possam ser devidamente fornecidos a determinados algoritmos de Aprendizado de Máquina.

- **Normalização:** por fim, a normalização é utilizada para ajustar a escala das características numéricas em um conjunto de dados. O objetivo é transformar as características para que fiquem em uma escala comum, geralmente para melhorar a performance e a estabilidade dos algoritmos de Aprendizado de Máquina.

3 DEFINIÇÃO DA ABORDAGEM, MÉTRICAS E MÉTODOS DE AVALIAÇÃO

A abordagem a ser utilizada consiste na utilização de algoritmos de aprendizado supervisionado, a fim de realizar predições a respeito do atributo Tipo do Clima (*Weather Type*), conforme mencionado na Seção 2.

Para utilização dos algoritmos, foi utilizada a biblioteca *scikit-learn*, em linguagem Python. Foram utilizados três algoritmos diferentes sobre o conjunto dos dados, sendo eles: Árvore de Decisão, k-Nearest Neighbors (k-NN) e Naive Bayes. O algoritmo k-Nearest Neighbors, especificamente, foi utilizado com três diferentes cálculos de distância: Distância Manhattan, Distância Euclidiana e Distância de Chebyshev. Assim, considera-se que existem um total de cinco modelos diferentes incluídos no processo de testes. A métrica utilizada, para aferir o quão completo e genérico os modelos estão, consiste na acurácia. Para obtenção das curvas PR e ROC durante a elaboração da Etapa 2 do projeto, também foram consideradas as métricas da taxa de verdadeiros positivos, taxa de falsos positivos, precisão e *recall*. Os métodos de avaliação incluem a utilização de *holdout* e *bootstrap*, conforme explicado nas Seções 4 e 5.

4 SPOT-CHECKING

Após o pré-processamento dos dados mencionados na Seção 2, foram aplicados os algoritmos indicados na Seção 3. A Tabela 1 indica a média aritmética das acurácias obtidas individualmente para cada um dos cinco modelos, ao longo de dez execuções diferentes. Essa tabela foi obtida durante a elaboração da Etapa 1, com a utilização do método *holdout* na proporção 85/15, permitindo uma visualização preliminar dos algoritmos mais promissores.

Tabela 1 - Média aritmética das acurácias individuais obtidas após dez execuções diferentes, para cada modelo

Modelo	Média
Árvore de Decisão	0,9708
k-NN com Distância Manhattan	0,9604
k-NN com Distância Euclidiana	0,9584
k-NN com Distância de Chebyshev	0,9426
Naive Bayes	0,8827

A Figura 1 permite visualizar graficamente, por meio de um diagrama de colunas, as diferenças de acurácias obtidas ao final do *Spot-checking*.

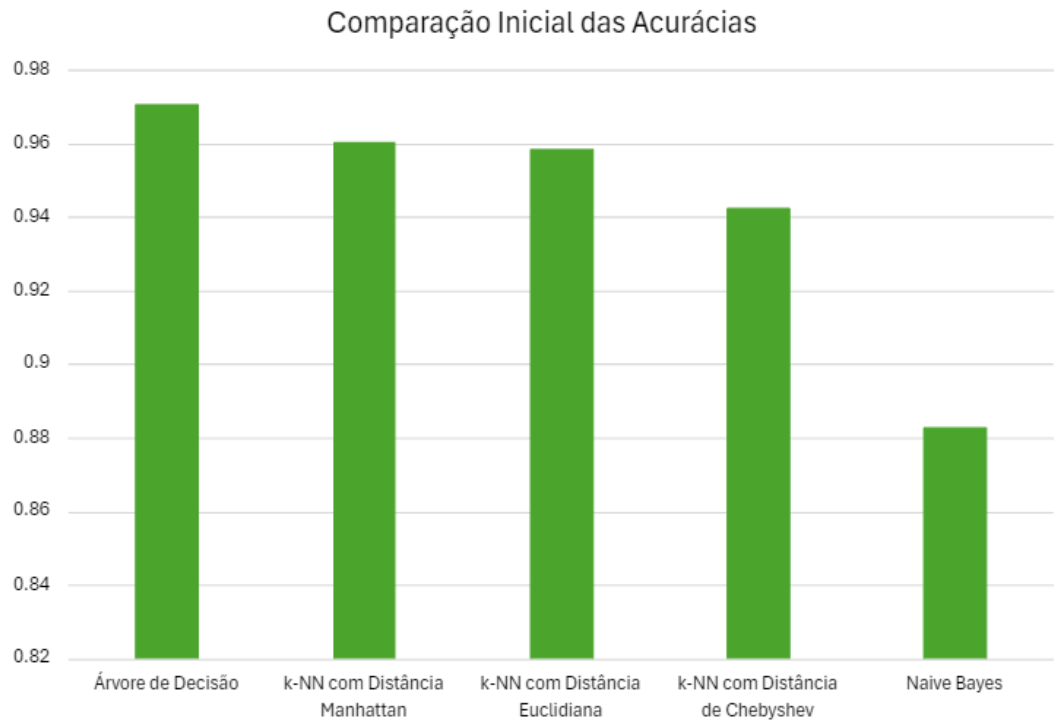


Figura 1 - Comparação Gráfica das Acurácias Obtidas no *Spot-checking*

As acurácias indicadas na Tabela 1 e na Figura 1 revelaram, inicialmente, uma alta acurácia das predições dos modelos, sobretudo para os algoritmos de Árvore de Decisão e k-NN.

5 TREINAMENTO E AVALIAÇÃO DOS MODELOS

Na Etapa 2 do desenvolvimento do projeto, os conjuntos de dados de treinamento, validação e teste foram divididos de tal forma que 70% das instâncias sejam destinadas para treinamento, 15% das instâncias sejam destinadas para validação e 15% das instâncias sejam destinadas para testes. Em cada execução, as instâncias selecionadas para compor cada um dos três conjuntos são escolhidas de forma aleatória, com a observância de que deve existir sempre o mesmo percentual de instâncias de cada classe (*cloudy*, *snowy*, *sunny* ou *rainy*) em cada um dos três conjuntos. Em seguida, são aplicadas as etapas de pré-processamento de dados, descritas na Seção 2. Para avaliação dos modelos, foi realizada a otimização de hiperparâmetros com o conjunto de validação.

Realizou-se o cálculo da acurácia para cada um dos modelos, de forma individual, a fim de determinar a capacidade de predição. Foram aplicados os métodos de *holdout* e de *bootstrap* para avaliação dos modelos com base na acurácia, também permitindo comparar os valores da acurácia obtidos para cada um dos dois métodos de avaliação. Adicionalmente, também foram medidos os impactos gerados nos casos em que não ocorre a remoção de *outliers*.

5.1 Testes envolvendo a Árvore de Decisão

Para os testes envolvendo a Árvore de Decisão, foi realizada a otimização do hiperparâmetro relacionado com a profundidade máxima da árvore, com valores de teste de profundidade variando entre 1 e 100 durante a análise envolvendo o conjunto de validação. A Tabela 2 e o gráfico de colunas da Figura 2 mostram os valores da acurácia obtidos em quatro execuções diferentes, variando entre a utilização do método *holdout* ou do método de *bootstrap*, também alternando entre remover ou não remover os *outliers* do conjunto de dados. Para cada execução, é indicado o valor do hiperparâmetro relacionado com a profundidade máxima da árvore.

Tabela 2 - Acurácias com Árvore de Decisão

Método de Avaliação	Presença de *Outliers*	Profundidade Máxima	Acurácia Obtida
Holdout	Sim	8	0,906565657
Holdout	Não	10	0,971806674
Bootstrap	Sim	9	0,926262626
Bootstrap	Não	38	0,963793103

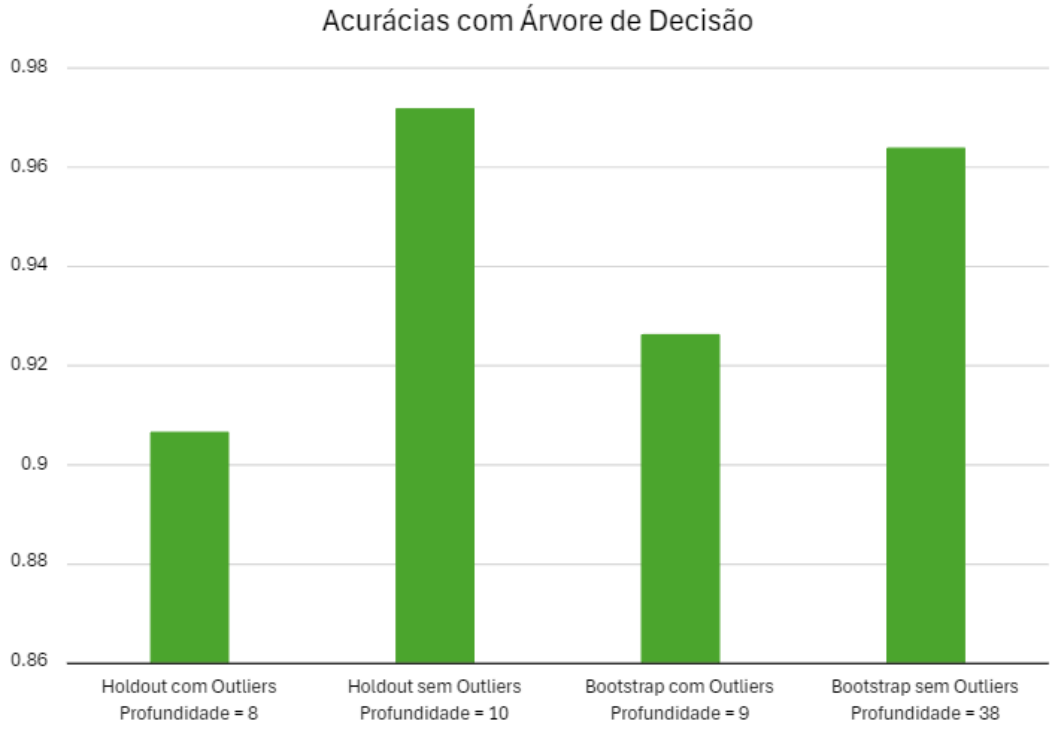


Figura 2 - Comparação Gráfica das Acurácias Obtidas com Árvore de Decisão

Abaixo, são apresentadas as curvas ROC e PR para o caso específico da execução com maior acurácia, na qual foi utilizado o método *holdout* com eliminação de *outliers*, conforme indicado na Tabela 2. Para as curvas ROC, são apresentados os casos envolvendo cada um dos possíveis valores do atributo alvo (*cloudy*, *rainy, snowy* ou *sunny*).

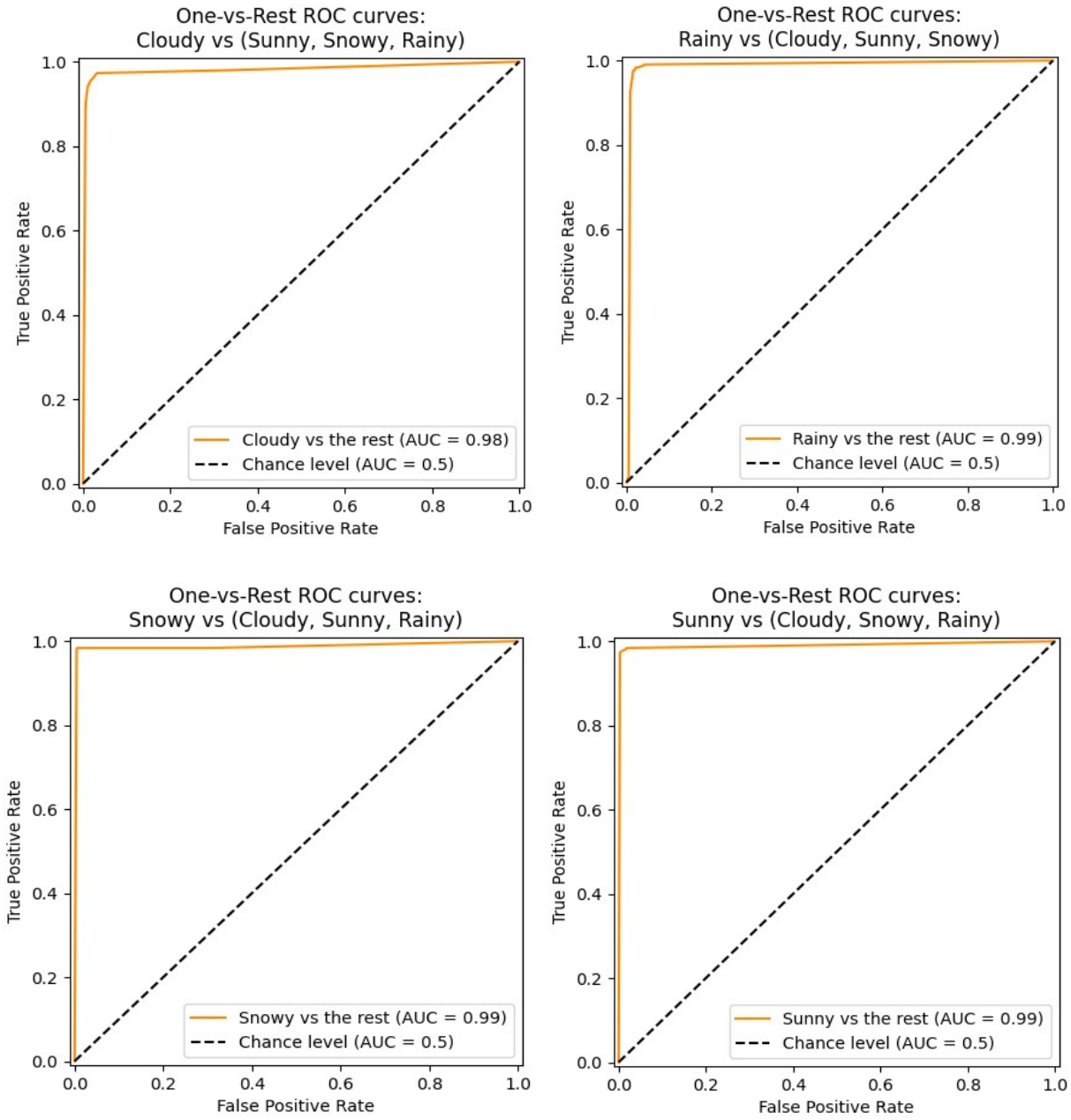


Figura 3 - Curvas ROC das Classificações Envolvendo Árvore de Decisão

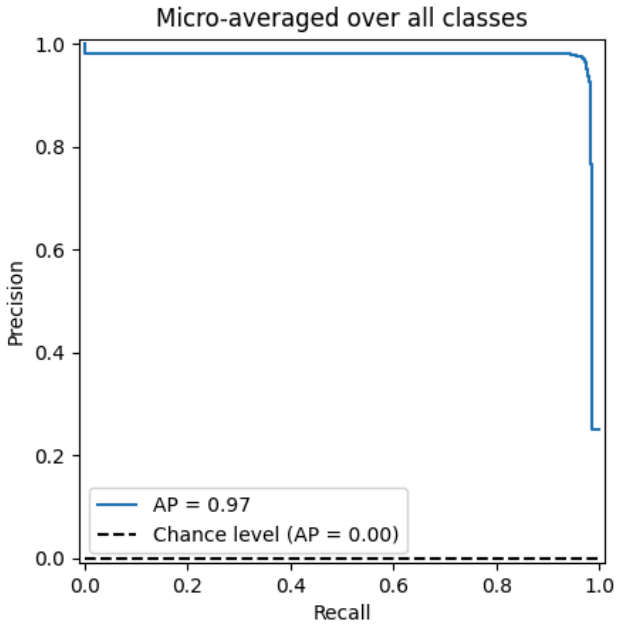


Figura 4 - Curva PR das Classificações Envolvendo Árvore de Decisão

5.2 Testes envolvendo o algoritmo k-NN

Para os testes envolvendo os modelos com o algoritmo k-NN, para cada um dos três métodos do cálculo de distância selecionados, são testados todos os valores ímpares variando de 1 até 11 para o hiperparâmetro *k*, levando em consideração as instâncias do conjunto de validação, a fim de otimizar o hiperparâmetro. Por fim, é retornada a acurácia obtida com o melhor valor de *k* selecionado.

5.2.1 Testes envolvendo o algoritmo k-NN e Distância Euclidiana

A Tabela 3 e o gráfico de colunas da Figura 5 mostram os valores da acurácia obtidos em quatro execuções diferentes, variando entre a utilização do método *holdout* ou do método de *bootstrap*, também alternando entre remover ou não remover os *outliers* do conjunto de dados. Para cada execução, é indicado o valor do hiperparâmetro k do algoritmo k-NN.

Tabela 3 - Acurácias com k-NN e Distância Euclidiana			
Método de Avaliação	**Presença de *Outliers***	**Hiperparâmetro k**	**Acurácia Obtida**
Holdout	Sim	1	0,891919192
Holdout	Não	1	0,945914845
Bootstrap	Sim	7	0,893434343
Bootstrap	Não	1	0,959195402

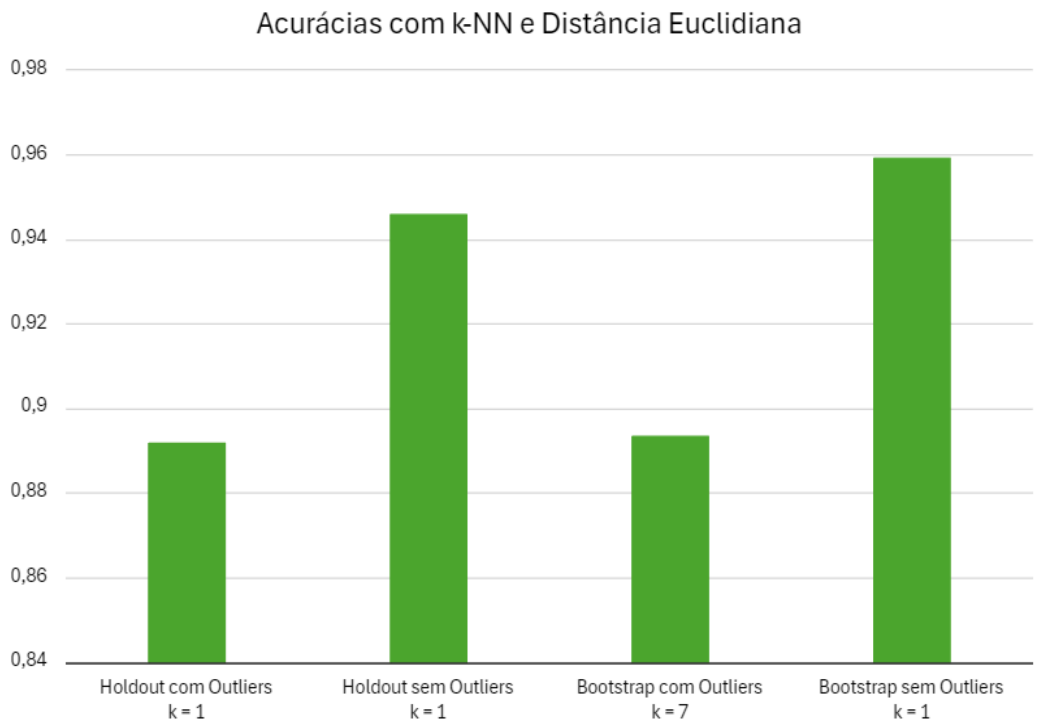


Figura 5 - Comparação Gráfica das Acurácias Obtidas com k-NN e Distância Euclidiana

Abaixo, são apresentadas as curvas ROC e PR para o caso específico da execução com maior acurácia, na qual foi utilizado o método *bootstrap* com eliminação de *outliers*, conforme indicado na Tabela 3. Para as curvas ROC, são apresentados os casos envolvendo cada um dos possíveis valores do atributo alvo (*cloudy*, *rainy, snowy* ou *sunny*).

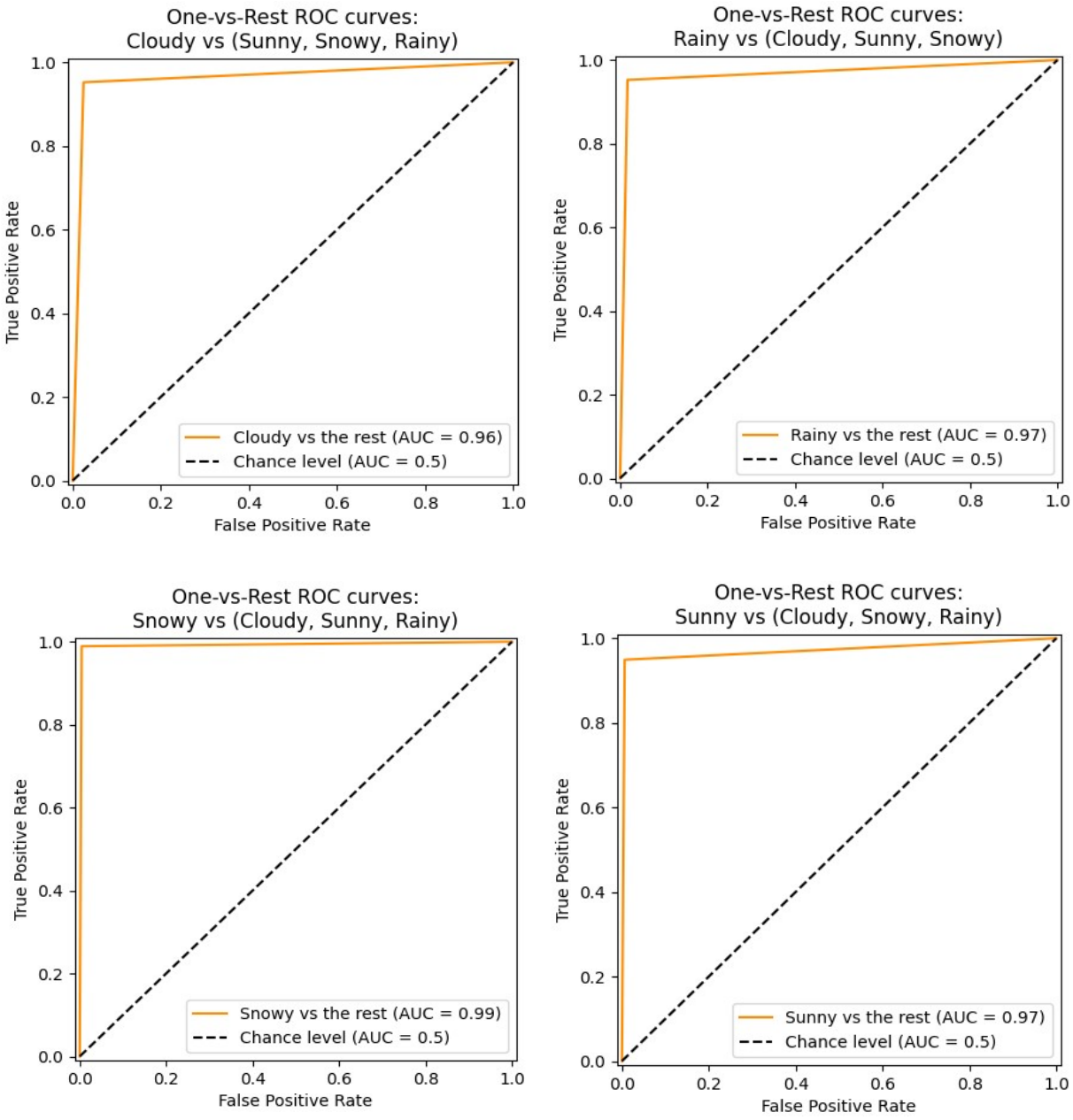


Figura 6 - Curvas ROC das Classificações Envolvendo k-NN com Distância Euclidiana

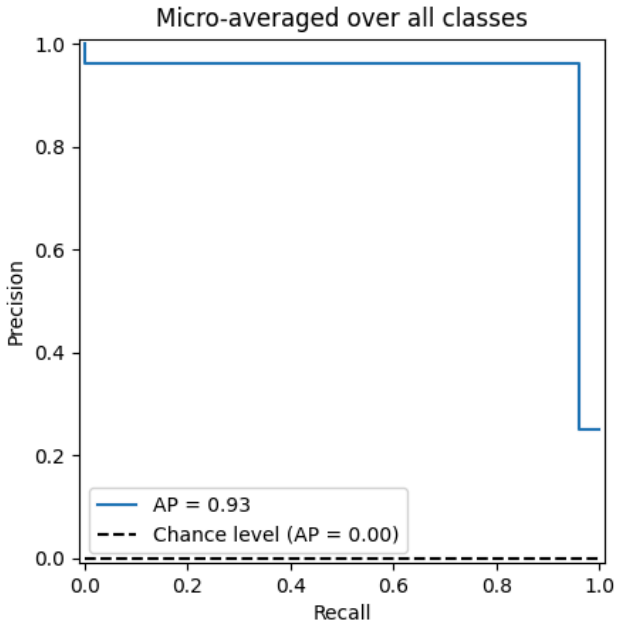


Figura 7 - Curva PR das Classificações Envolvendo k-NN com Distância Euclidiana

5.2.2 Testes envolvendo o algoritmo k-NN e Distância de Chebyshev

A Tabela 4 e o gráfico de colunas da Figura 8 mostram os valores da acurácia obtidos em quatro execuções diferentes, variando entre a utilização do método *holdout* ou do método de *bootstrap*, também alternando entre remover ou não remover os *outliers* do conjunto de dados. Para cada execução, é indicado o valor do hiperparâmetro k do algoritmo k-NN.

Tabela 4 - Acurácias com k-NN e Distância de Chebyshev			
Método de Avaliação	**Presença de *Outliers***	**Hiperparâmetro k**	**Acurácia Obtida**
Holdout	Sim	1	0,887373737
Holdout	Não	1	0,943613349
Bootstrap	Sim	5	0,896464646
Bootstrap	Não	7	0,946551724

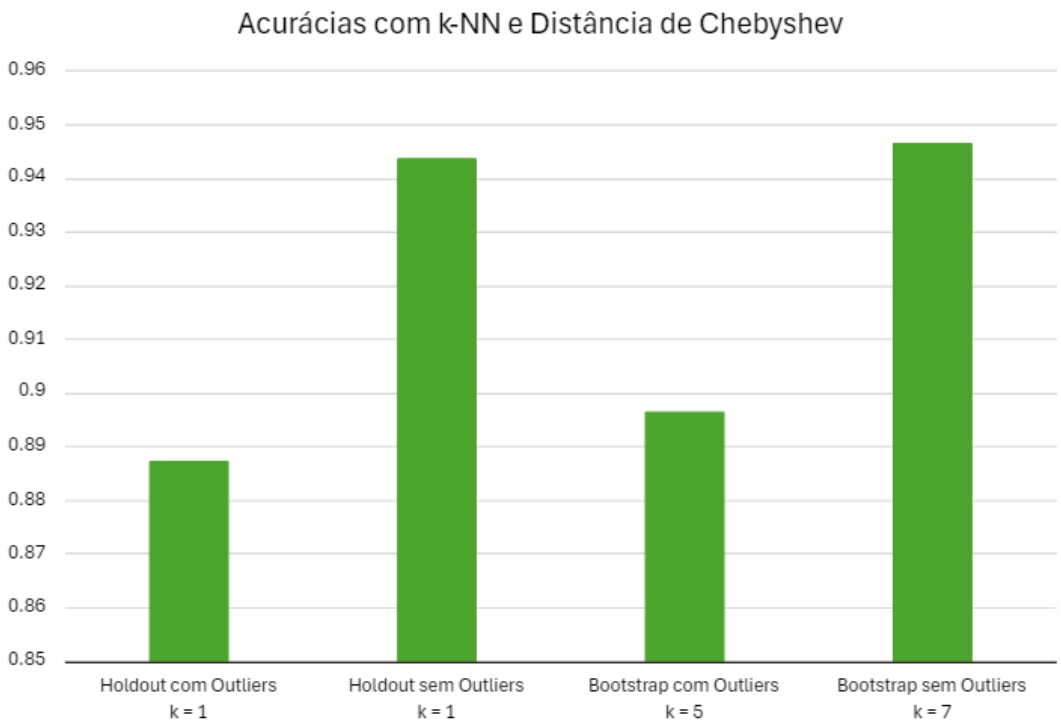


Figura 8 - Comparação Gráfica das Acurácias Obtidas com k-NN e Distância de Chebyshev

Abaixo, são apresentadas as curvas ROC e PR para o caso específico da execução com maior acurácia, na qual foi utilizado o método *bootstrap* com eliminação de *outliers*, conforme indicado na Tabela 4. Para as curvas ROC, são apresentados os casos envolvendo cada um dos possíveis valores do atributo alvo (*cloudy*, *rainy, snowy* ou *sunny*).

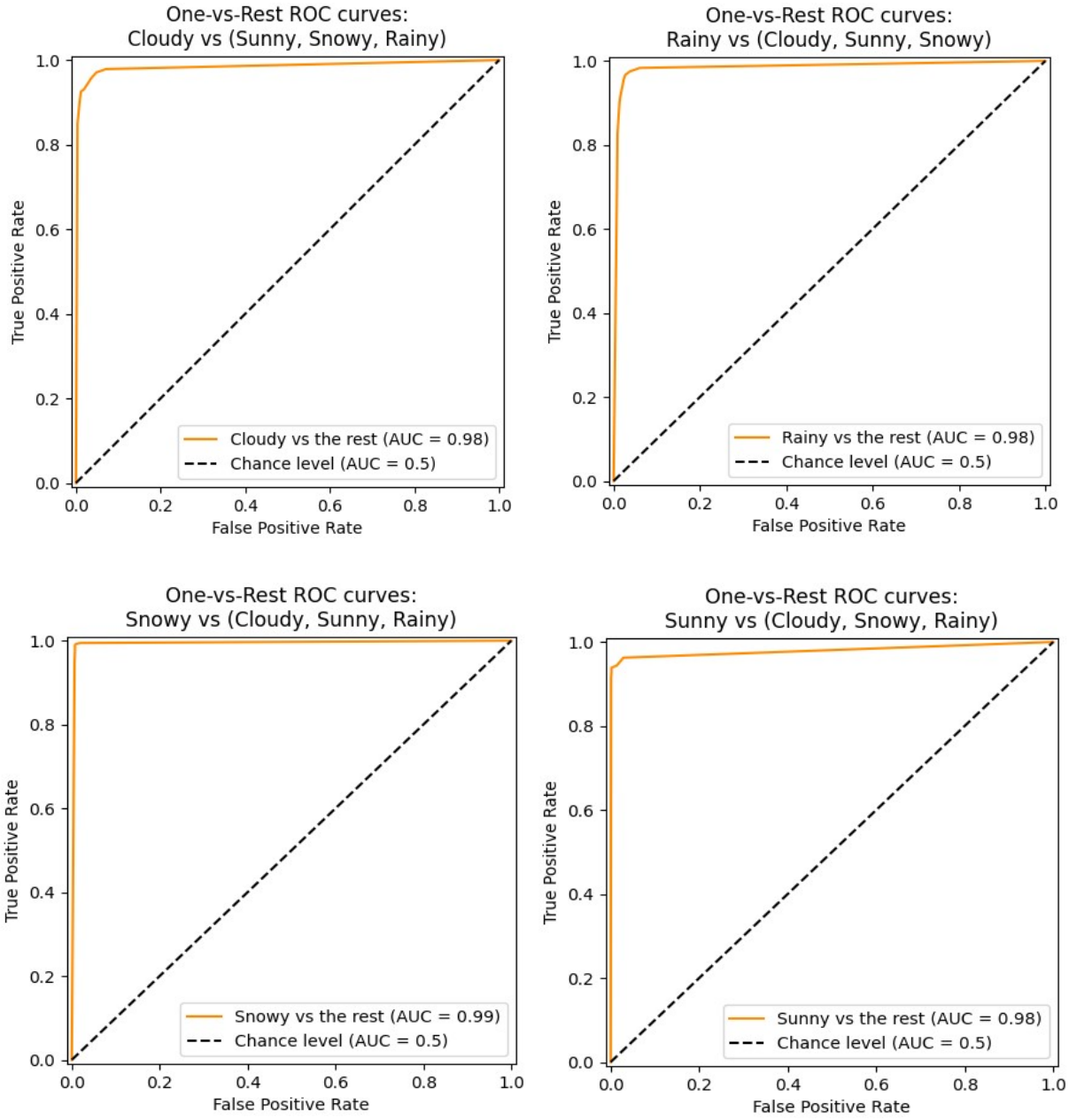


Figura 9 - Curvas ROC das Classificações Envolvendo k-NN com Distância de Chebyshev

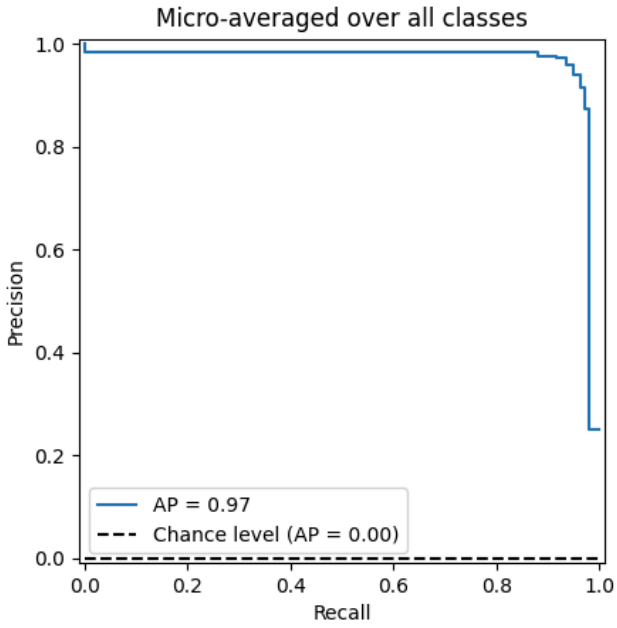


Figura 10 - Curva PR das Classificações Envolvendo k-NN com Distância de Chebyshev

5.2.3 Testes envolvendo o algoritmo k-NN e Distância Manhattan

A Tabela 5 e o gráfico de colunas da Figura 11 mostram os valores da acurácia obtidos em quatro execuções diferentes, variando entre a utilização do método *holdout* ou do método de *bootstrap*, também alternando entre remover ou não remover os *outliers* do conjunto de dados. Para cada execução, é indicado o valor do hiperparâmetro k do algoritmo k-NN.

Tabela 5 - Acurácias com k-NN e Distância Manhattan

Método de Avaliação	**Presença de *Outliers***	**Hiperparâmetro k**	**Acurácia Obtida**
Holdout	Sim	1	0,885353535
Holdout	Não	1	0,949367089
Bootstrap	Sim	9	0,895454545
Bootstrap	Não	5	0,951724138

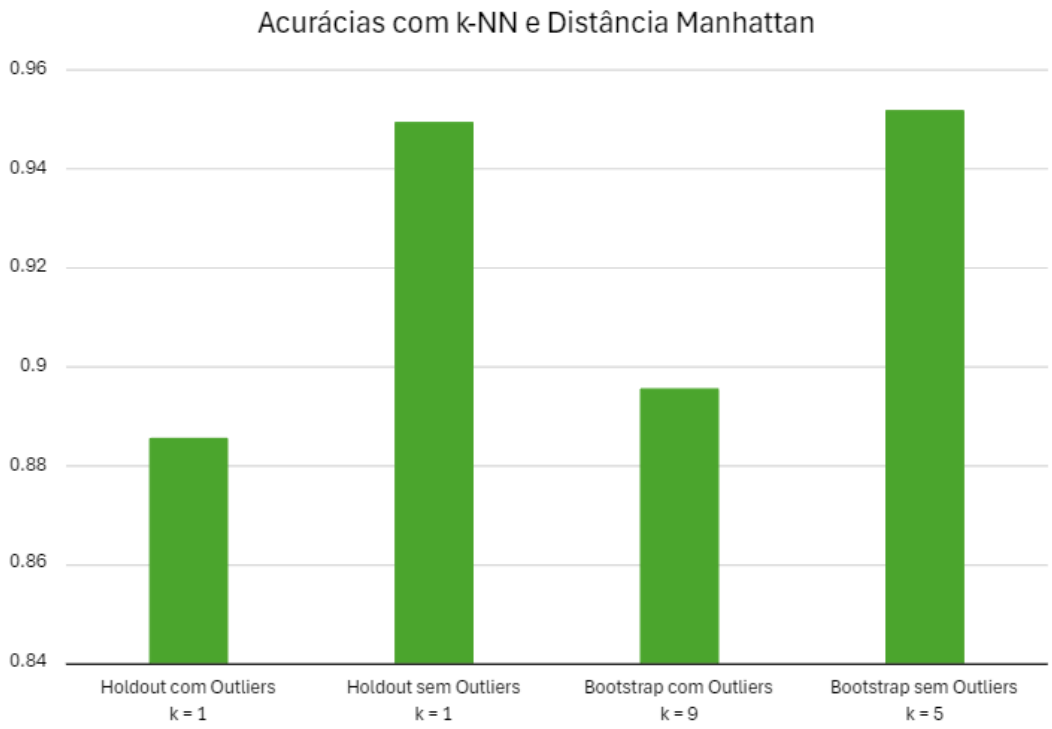


Figura 11 - Comparação Gráfica das Acurácias Obtidas com k-NN e Distância Manhattan

Abaixo, são apresentadas as curvas ROC e PR para o caso específico da execução com maior acurácia, na qual foi utilizado o método *bootstrap* com eliminação de *outliers*, conforme indicado na Tabela 5. Para as curvas ROC, são apresentados os casos envolvendo cada um dos possíveis valores do atributo alvo (*cloudy*, *rainy, snowy* ou *sunny*).

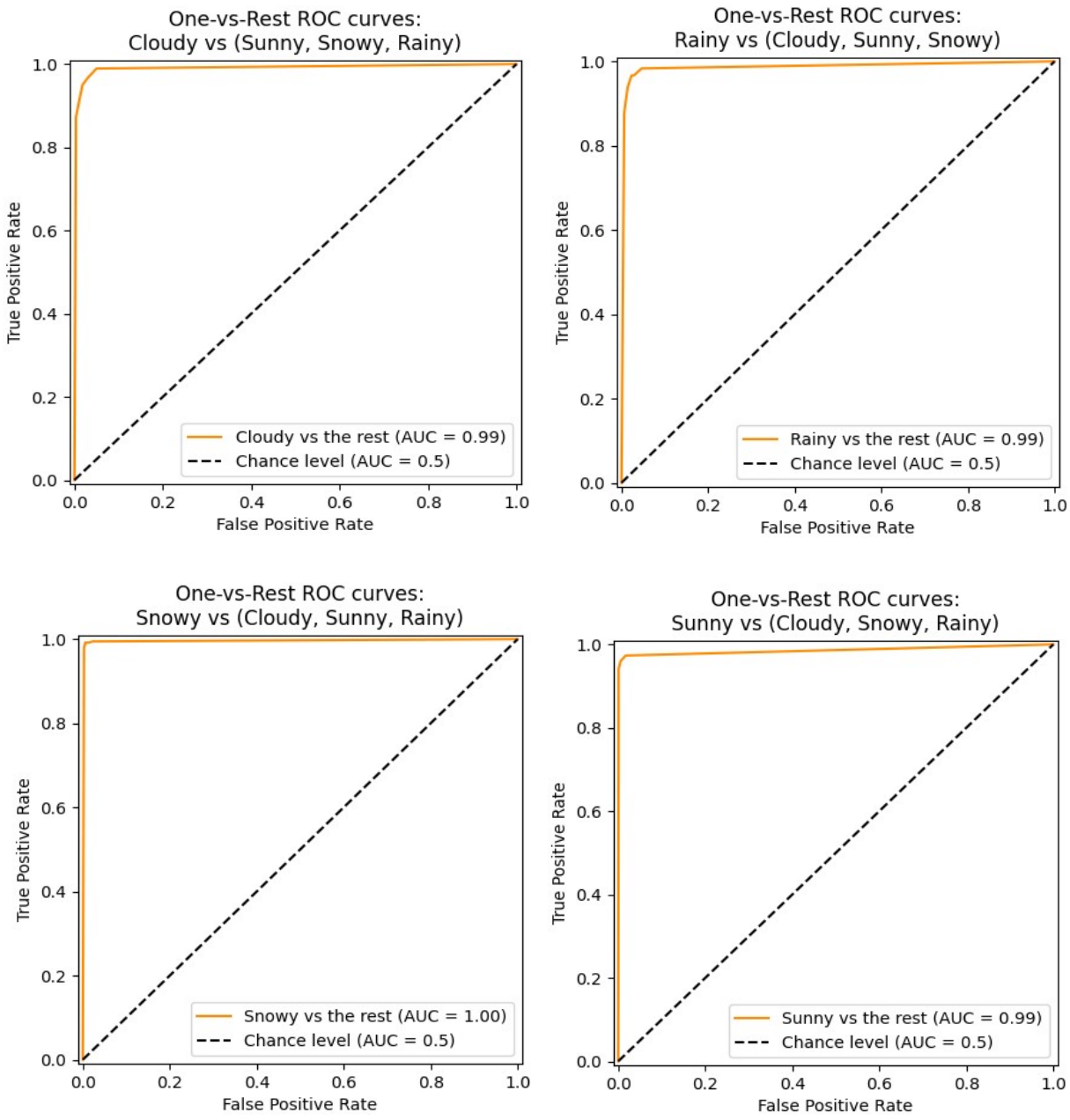


Figura 12 - Curvas ROC das Classificações Envolvendo k-NN com Distância Manhattan

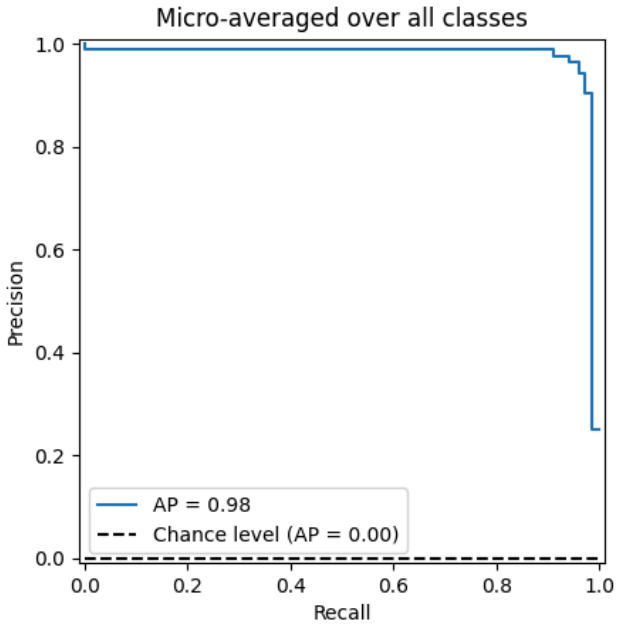


Figura 13 - Curva PR das Classificações Envolvendo k-NN com Distância Manhattan

5.3 Testes envolvendo o algoritmo Naive Bayes

Nos testes envolvendo o modelo Naive Bayes, ocorre a aplicação do Naive Bayes Gaussiano para análise dos atributos numéricos, e a aplicação do Naive Bayes Multinomial para análise dos atributos categóricos, nas instâncias a serem testadas. Os resultados obtidos com as duas variações são combinados por meio do produto dos valores de predição obtidos. Para o caso específico desse modelo, não há otimização de hiperparâmetros envolvida.

A Tabela 6 e o gráfico de colunas da Figura 14 mostram os valores da acurácia obtidos em quatro execuções diferentes, variando entre a utilização do método *holdout* ou do método de *bootstrap*, também alternando entre remover ou não remover os *outliers* do conjunto de dados.

Tabela 6 - Acurácias com Naive Bayes

Método de Avaliação	**Presença de *Outliers***	**Acurácia Obtida**
Holdout	Sim	0,768686869
Holdout	Não	0,867663982
Bootstrap	Sim	0,774747475
Bootstrap	Não	0,886206897

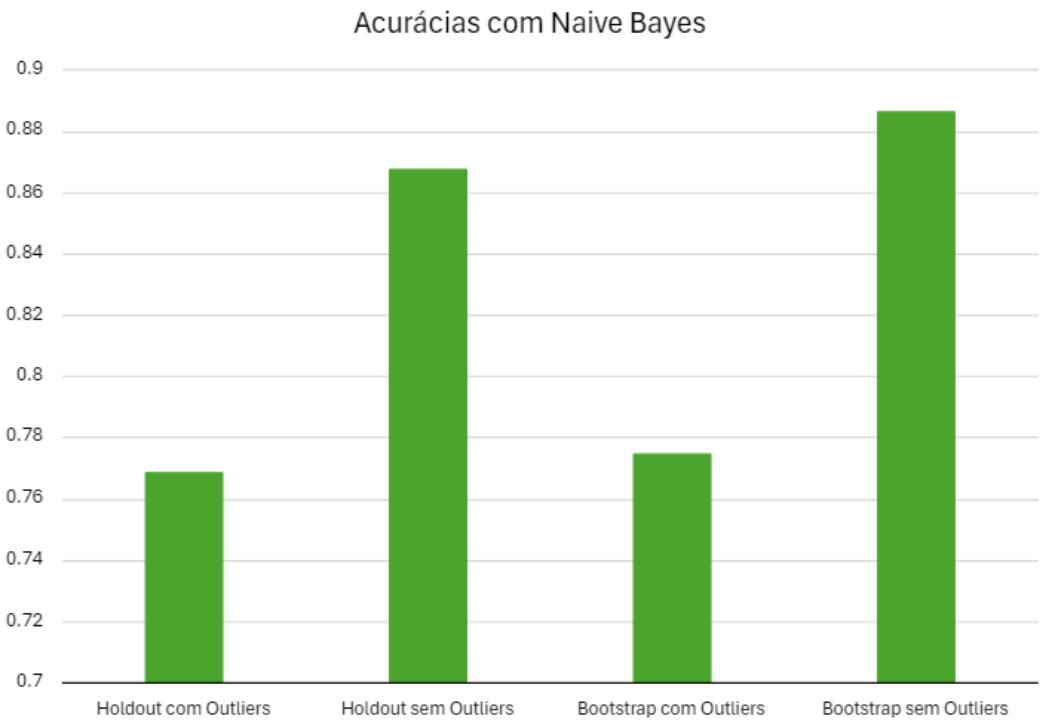
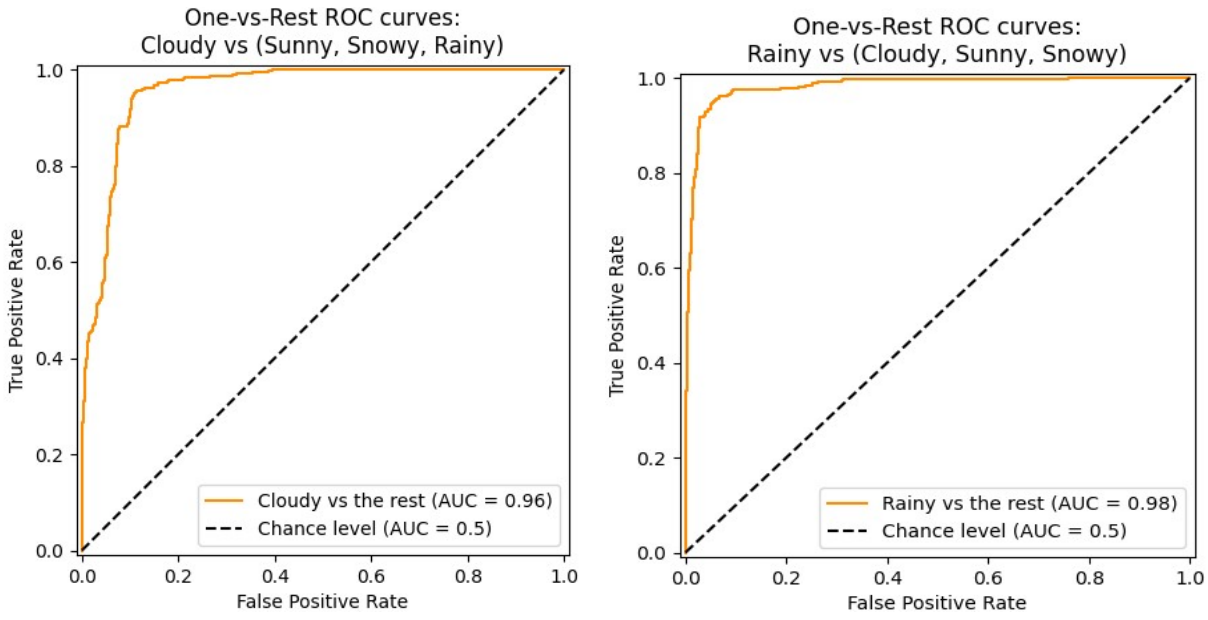


Figura 14 - Comparação Gráfica das Acurácias Obtidas com k-NN e Distância Manhattan

Abaixo, são apresentadas as curvas ROC e PR para o caso específico da execução com maior acurácia, na qual foi utilizado o método *bootstrap* com eliminação de *outliers*, conforme indicado na Tabela 6. Para as curvas ROC, são apresentados os casos envolvendo cada um dos possíveis valores do atributo alvo (*cloudy*, *rainy, snowy* ou *sunny*).



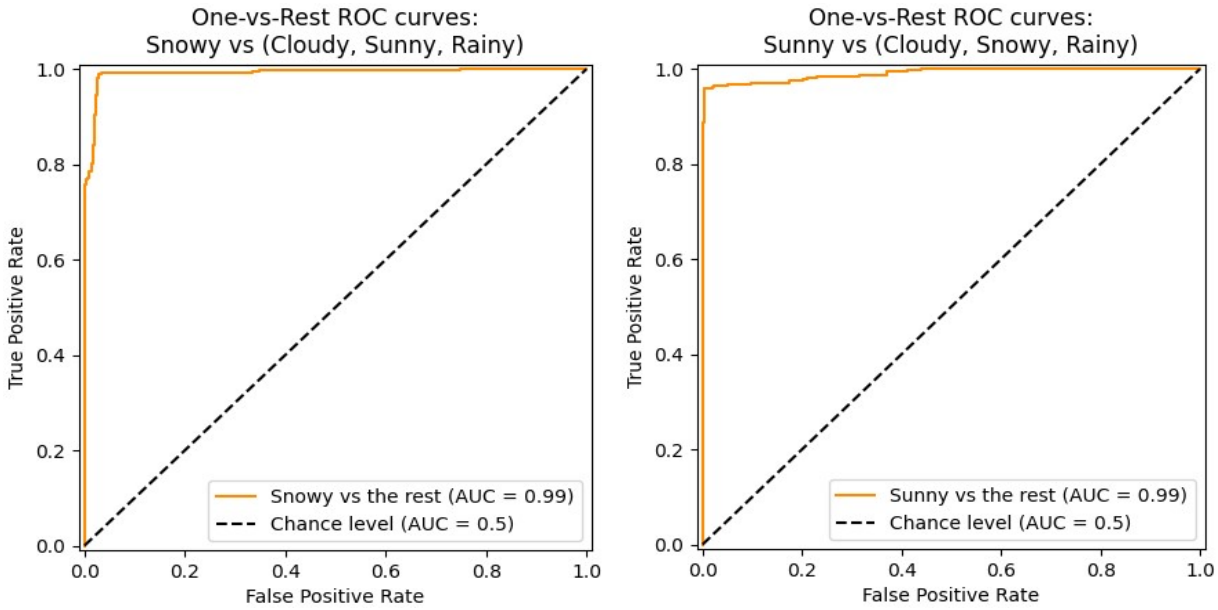


Figura 15 - Curvas ROC das Classificações Envolvendo Naive Bayes

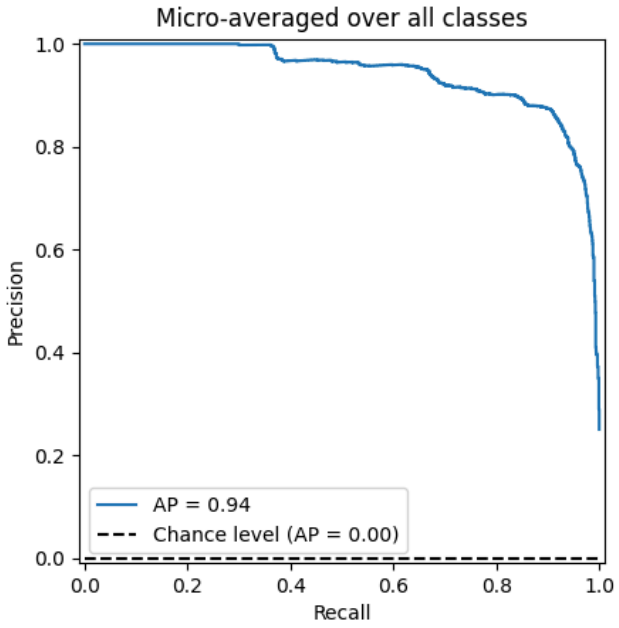


Figura 16 - Curva PR das Classificações Envolvendo Naive Bayes

6 COMPARAÇÃO DOS MODELOS

A Tabela 7 e a Figura 17 apresentam uma comparação envolvendo os cinco modelos, com o método de avaliação *holdout*.

Tabela 7 - Comparação Numérica das Melhores Acurácias Obtidas com os Modelos, com Método *holdout*

Algoritmo	Presença de *Outliers*	Valor do Hiperparâmetro	Acurácia
Árvore de Decisão	Não	10	0,971806674
k-NN com Distância Euclidiana	Não	1	0,945914845
k-NN com Distância Manhattan	Não	1	0,949367089
k-NN com Distância de Chebyshev	Não	1	0,943613349
Naive Bayes	Não	-	0,867663982

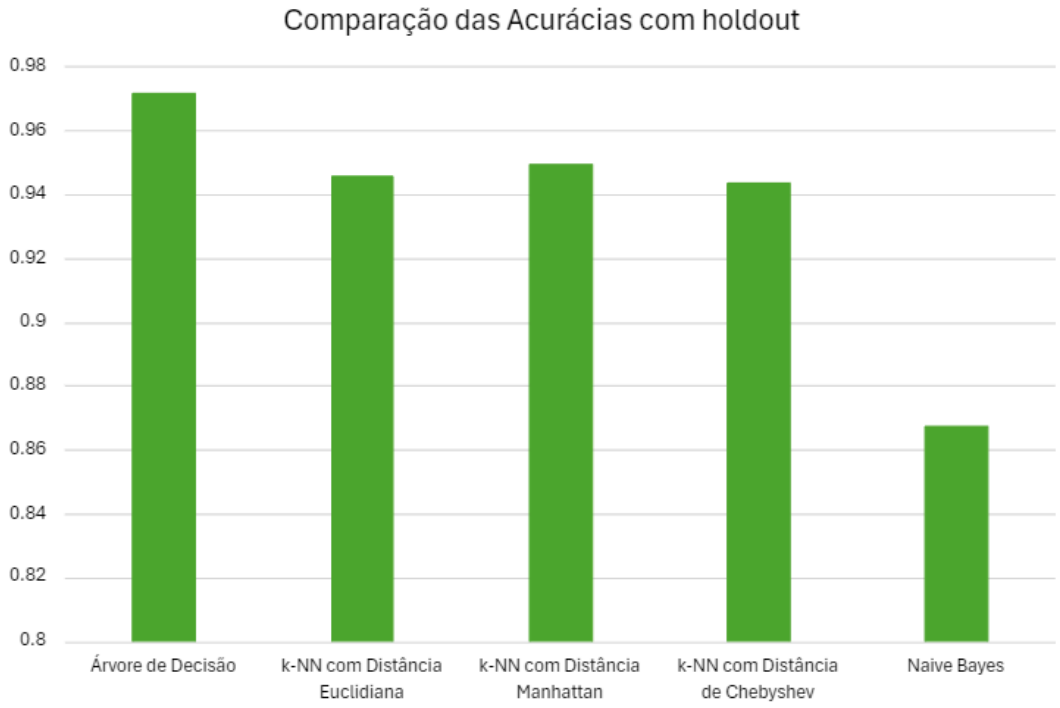


Figura 17 - Comparação Gráfica das Melhores Acurácias Obtidas com *holdout*

A Tabela 8 e a Figura 18 apresentam uma outra comparação envolvendo os cinco modelos, com o método de avaliação *bootstrap*.

Tabela 8 - Comparação Numérica das Melhores Acurácias Obtidas com os Modelos, com Método *bootstrap*			
Algoritmo	**Presença de *Outliers***	**Valor do Hiperparâmetro**	**Acurácia**
Árvore de Decisão	Não	38	0,963793103
k-NN com Distância Euclidiana	Não	1	0,959195402
k-NN com Distância Manhattan	Não	5	0,951724138
k-NN com Distância de Chebyshev	Não	7	0,946551724
Naive Bayes	Não	-	0,886206897

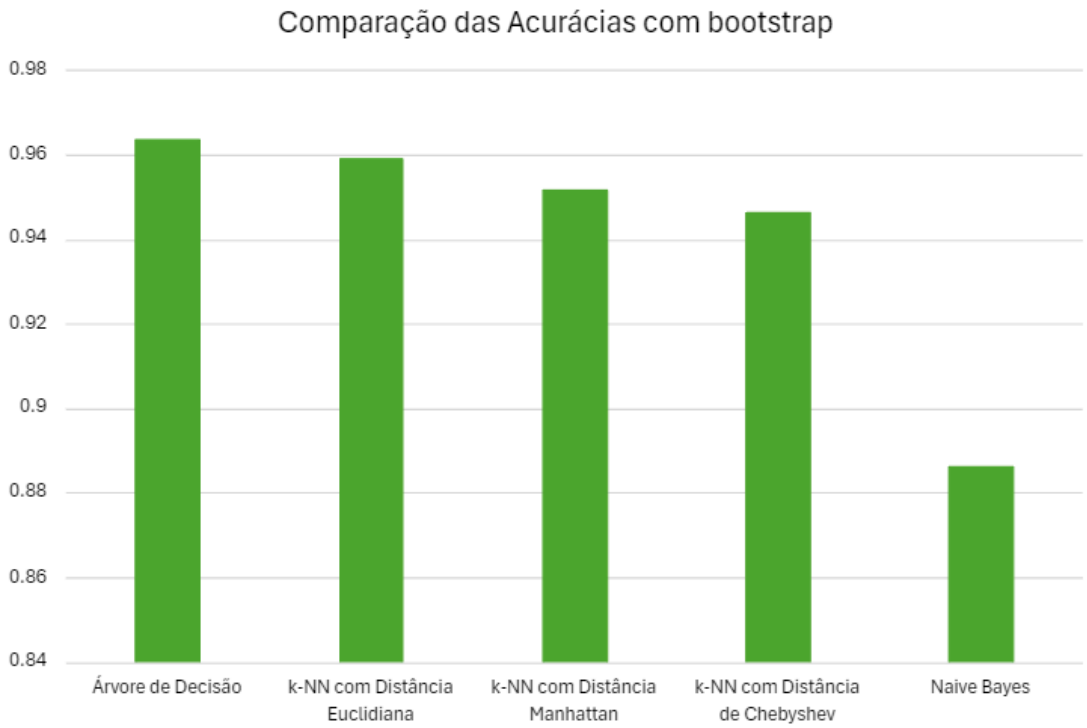


Figura 18 - Comparação Gráfica das Melhores Acurácias Obtidas com *bootstrap*

A Tabela 9 apresenta um resumo com as melhores acurácias obtidas para qualquer um dos dois métodos de avaliação. Em todos os casos, a remoção de *outliers* contribuiu para um aumento na acurácia.

Tabela 9 - Comparação Numérica das Melhores Acurácias Obtidas com os Modelos

Algoritmo	Método de Avaliação	Presença de *Outliers*	Valor do Hiperparâmetro	Acurácia
Árvore de Decisão	*Holdout*	Não	10	0,971806674
k-NN com Distância Euclidiana	*Bootstrap*	Não	1	0,959195402
k-NN com Distância Manhattan	*Bootstrap*	Não	5	0,951724138
k-NN com Distância de Chebyshev	*Bootstrap*	Não	7	0,946551724
Naive Bayes	*Bootstrap*	Não	-	0,886206897

O gráfico de colunas da Figura 19 permite visualizar e comparar de forma mais direta o melhor valor de acurácia obtido, para cada um dos cinco modelos considerados.

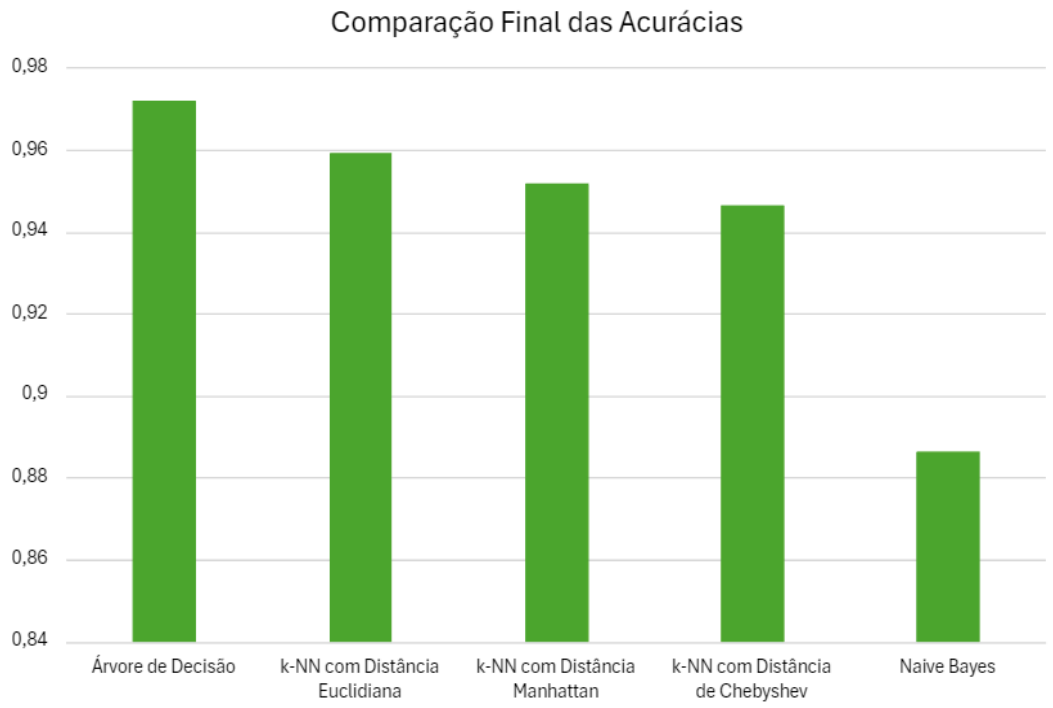


Figura 19 - Comparação Gráfica das Melhores Acurácias Obtidas

7 CONCLUSÃO

No desenvolvimento da Etapa 1 do projeto, o *Spot-checking* havia revelado, de forma preliminar, uma superioridade do modelo de Árvore de Decisão em relação aos demais modelos. Os resultados expostos nas Seções 5 e 6 revelam que essa superioridade foi predominante na Etapa 2 do desenvolvimento do projeto, na qual houve alteração na forma de otimização de hiperparâmetros (utilizando um conjunto de validação especificamente para essa finalidade), estratificação na divisão dos dados (fazendo com que exista o mesmo percentual de cada classe em cada um dos três conjuntos), variação no método de avaliação (*holdout* e *bootstrap*) e análise do impacto da presença e da remoção de *outliers*.

O *Spot-checking* também revelou valores de acurácias altos e semelhantes para os modelos envolvendo k-NN, sobretudo com as distâncias Euclidiana e Manhattan, que apresentaram valores de acurácia bastante semelhantes, embora o modelo envolvendo a distância de Chebyshev também tenha gerado bons resultados. Conforme também revelado de forma preliminar, o modelo Naive Bayes demonstrou o pior resultado entre todos os cinco modelos durante a elaboração de ambas as etapas, mesmo que a remoção de *outliers* também tenha contribuído para aumentos significativos na acurácia desse modelo em especial.

Além dos pré-processamentos essenciais, que foram mantidos durante todos os testes realizados na Etapa 2 (normalização e *one-hot-encoding*), a realização desses testes evidenciou a importância da remoção dos dados ruidosos do conjunto de dados em análise, visto que esse pré-processamento dos dados gerou impacto significativo nos valores de acurácia para todos os modelos.

O método de avaliação de *bootstrap*, introduzido na Etapa 2, foi capaz de revelar uma alta acurácia para a maioria dos modelos testados, revelando uma robustez existente nos modelos, uma vez que esse método permite a obtenção de uma média de desempenho do modelo com várias amostras, fornecendo uma avaliação mais robusta em relação a uma simples divisão treino-teste. Entretanto, o método de avaliação de *holdout* apresentou valores semelhantes aos valores obtidos com *bootstrap* para cada um dos cinco modelos, com a remoção de *outliers* impactando mais nos valores de acurácia, em comparação com o método de avaliação.

Possibilidades de continuação e extensão dos trabalhos desenvolvidos nas Etapas 1 e 2 podem incluir:

- Criação de ensembles envolvendo os modelos desenvolvidos;
- Análise de desempenho dos modelos utilizando outros métodos de avaliação, como *cross validation*;

- Análises envolvendo outros hiperparâmetros para a Árvore de Decisão, como o número mínimo de amostras para dividir um nó, ou o número mínimo de amostras que um nó folha deve possuir;
- Adição de novos modelos para a tarefa de classificação (envolvendo, por exemplo, o uso de redes neurais), de modo a tentar obter um novo modelo com acurácia muito próxima ou superior à melhor acurácia obtida com os modelos já existentes;
- Análise de desempenho, mediante a aplicação dos modelos desenvolvidos, em outros conjuntos de dados, relacionados com um problema de classificação semelhante.