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 I

Disciplina: Aprendizado de Máquina

Professor: Anderson Rocha Tavares

Porto Alegre, julho 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 “primavera”, “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.

Inicialmente, o conjunto de dados de treinamento e o conjunto de dados de teste foi dividido na proporção 85/15, para utilização do método *holdout*. Dessa forma, as instâncias usadas para treinamento e as instâncias utilizadas para teste são escolhidas de forma aleatória, em cada execução. Em seguida, são aplicadas as etapas de pré-processamento de dados, descritas na Seção 2.

A métrica utilizada para aferir o quão completo e genérico os modelos estão consiste na acurácia. Realizou-se o cálculo da acurácia para cada um dos modelos, de forma individual, a fim de determinar quais são os algoritmos mais promissores.

4 SPOT-CHECKING

Após o pré-processamento dos dados, 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.

Nos testes envolvendo o modelo 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*, com o cálculo da acurácia sendo realizado para cada um desses casos. Por fim, é retornada a acurácia obtida com o melhor valor de *k* selecionado.

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.

Os testes envolvendo a Árvore de Decisão não incluíram procedimentos de poda da árvore.

A Tabela 1 indica a média aritmética das acurácias obtidas individualmente para cada um dos cinco modelos, ao longo de dez execuções diferentes.

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

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