

Lista 4

▼ Questão 1

Questão 1.

Tabela de probabilidade:

Jogar	Aparência			Temperatura			Umidade		Ventando	
	sol	nublado	chuva	quente	agrad.	fria	alta	normal	S	N
Não	$\frac{3}{5}$	$\frac{0}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{2}{5}$
Sim	$\frac{1}{9}$	$\frac{4}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{6}{9}$	$\frac{3}{9}$	$\frac{6}{9}$

$Jogar = \frac{9}{14} \times \frac{3}{5} \times \frac{3}{9} \times \frac{6}{9} \times \frac{3}{9} = 0.0182 / 0.0192 = 0.823 = 82.3\%$
 $Não Jogar = \frac{5}{14} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{3}{5} = 0.0034 / 0.0192 = 0.177 = 17.7\%$
 $Soma = 0.0192$

▼ Questão 3

1. Complexidade do Modelo:

- **Random Forest:** O Random Forest é um modelo de conjunto que combina várias árvores de decisão. Cada árvore é um modelo relativamente complexo, e o conjunto pode ser ainda mais complexo.
- **Naive Bayes:** O Naive Bayes é um modelo probabilístico simples que assume independência entre as features. É um modelo mais simples em comparação com o Random Forest.

2. Desempenho em Dados Categóricos:

- **Random Forest:** O Random Forest é capaz de lidar com variáveis categóricas após a codificação adequada (como no código enviado).
- **Naive Bayes :** O Naive Bayes é adequado para dados contínuos, mas pode não funcionar tão bem com variáveis categóricas sem algum tipo de transformação adicional.

3. Interpretabilidade:

- **Random Forest:** Os Random Forests são mais difíceis de interpretar devido à complexidade do conjunto de árvores. Você pode obter informações sobre a importância das features, mas a interpretação de decisões individuais pode ser desafiadora.
- **Naive Bayes:** O Naive Bayes é mais simples e interpretação-friendly, pois é baseado em probabilidades condicionais. É mais fácil entender como as features afetam as probabilidades de classificação.

4. Requisitos de Dados:

- **Random Forest:** O Random Forest pode lidar com uma ampla variedade de tipos de dados, mas é mais eficaz quando há muitas features.
- **Naive Bayes:** O Naive Bayes assume que as features são independentes e segue uma distribuição normal.

5. Hiperparâmetros:

- **Random Forest:** O Random Forest tem vários hiperparâmetros, como o número de estimadores, profundidade máxima da árvore, etc., que precisam ser ajustados para otimizar o desempenho.
- **Naive Bayes Gaussiano:** O Naive Bayes Gaussiano tem menos hiperparâmetros a serem ajustados, tornando-o mais fácil de configurar.

Em resumo, ambos os algoritmos têm suas vantagens e desvantagens. O Random Forest é mais poderoso e versátil, mas pode ser mais complexo de configurar. O Naive Bayes é mais simples e interpretação-friendly, adequado para conjuntos de dados menores e assume independência entre as features. A escolha entre eles depende dos requisitos específicos do seu problema e das características dos seus dados.

▼ Questão 4

1. RandomForestClassifier:

- **Técnica de Ensemble:** Random Forest é uma técnica de ensemble que combina múltiplas árvores de decisão para criar um modelo mais robusto e preciso.
- **Amostragem de Dados:** Usa bootstrap (amostragem com reposição)
- **Amostragem de Recursos:** Seleciona aleatoriamente um subconjunto de características em cada divisão de nó.
- **Votação:** As previsões de cada árvore são ponderadas igualmente ou por confiança e a classe mais votada é escolhida como a previsão final.

2. BaggingClassifier:

- **Técnica de Ensemble:** Bagging (Bootstrap Aggregating) é uma técnica de ensemble que cria múltiplos conjuntos de dados de treinamento usando amostragem com reposição e, em seguida, treina um modelo em cada conjunto.
- **Amostragem de Dados:** Usa bootstrap.
- **Amostragem de Recursos:** geralmente não faz isso.
- **Votação:** As previsões de cada modelo são combinadas por média ou voto (dependendo do problema).

3. GradientBoostingClassifier:

- **Técnica de Ensemble:** é uma técnica de ensemble que constrói uma série de modelos de maneira sequencial, onde cada modelo tenta corrigir os erros do modelo anterior.
- **Amostragem de Dados:** Não utiliza amostragem de dados; treina em todo o conjunto de treinamento em cada iteração.
- **Amostragem de Recursos:** não utiliza amostragem de recursos.
- **Aprendizado Sequencial:** A ênfase está na correção dos erros cometidos pelos modelos anteriores, tornando-o uma técnica de aprendizado sequencial.
- **Combinação de Modelos:** As previsões de cada modelo são combinadas com ponderação, onde os modelos mais precisos têm mais influência nas previsões finais.

Em resumo, a principal diferença entre essas técnicas de ensemble está na forma como elas constroem modelos e fazem previsões. O `RandomForestClassifier` cria várias árvores de decisão independentes e combina suas previsões, o `BaggingClassifier` cria vários modelos de maneira independente e os combina por votação ou média, enquanto o `GradientBoostingClassifier` constrói uma série de modelos sequencialmente, corrigindo os erros dos modelos anteriores.

▼ Questão 5

O artigo "A Survey of Ensemble Learning Concepts, Algorithms, Applications, and Prospects" realiza uma revisão abrangente dos conceitos, algoritmos, aplicações e perspectivas do aprendizado em conjunto. O estudo destaca a importância do aprendizado em conjunto como uma abordagem eficaz para melhorar o desempenho e a precisão dos modelos de aprendizado de máquina.

O artigo apresenta uma visão geral dos algoritmos de aprendizado em conjunto mais populares, como Random Forest, Bagging e Gradient Boosting. Descreve o Random Forest como uma técnica de ensemble que combina múltiplas árvores de decisão para criar um modelo mais robusto e preciso. O Bagging é descrito como uma técnica de ensemble que cria múltiplos conjuntos de dados de treinamento usando amostragem com reposição e, em seguida, treina um modelo em cada conjunto. Já o Gradient Boosting é apresentado como uma técnica de ensemble que constrói uma série de modelos sequencialmente, corrigindo os erros dos modelos anteriores.

O estudo destaca que essas técnicas de ensemble têm sido amplamente utilizadas em diversos domínios, como reconhecimento de padrões, detecção de fraudes, processamento de linguagem natural e bioinformática. Elas têm demonstrado melhorias significativas no desempenho preditivo, robustez e generalização dos modelos de aprendizado de máquina.

Em suma, o artigo enfoca os algoritmos de aprendizado de máquina utilizados no estudo, destacando o Random Forest, Bagging e Gradient Boosting como técnicas de ensemble eficazes para melhorar o desempenho e a precisão dos modelos de aprendizado de máquina.