UNIVERSIDADE FEDERAL DO TOCANTINS

CÂMPUS UNIVERSITÁRIO DE PALMAS

CURSO DE CIÊNCIA DA COMPUTAÇÃO

RELATÓRIO DE INTELIGÊNCIA ARTIFICIAL

KNN - APRENDIZADO SUPERVISIONADO

JOÃO GABRIEL ALVES DE SOUZA

JOÃO PEDRO SILVA CUNHA

PALMAS (TO)

2023

RESUMO

Este trabalho acadêmico apresenta uma visão geral do algoritmo KNN (k-nearest neighbors) no contexto do aprendizado supervisionado. O KNN é um dos métodos mais simples e populares para classificação e regressão, sendo amplamente utilizado em várias aplicações. O algoritmo se baseia no princípio de que objetos semelhantes tendem a estar próximos uns dos outros no espaço de características. O KNN classifica ou prevê a saída de um objeto desconhecido com base nas classes ou valores das amostras de treinamento mais próximas.Neste relatório, abordamos os principais conceitos e um breve experimento a cerca de duas bases de dados mostrando etapas de execução e considerações práticas ao utilizar o KNN e posteriormente comparar com os resultados submetidos ao WEKA usando as mesmas bases . Também discutimos suas vantagens e limitações, bem como algumas estratégias para melhorar seu desempenho. Por meio desta análise, o objetivo é fornecer uma compreensão clara do KNN e seu papel fundamental no campo do aprendizado supervisionado.

Palavra-chave: LaTeX. UFTeX. Relatório da Disciplina. KNN. Algoritmo de Aprendizado Supervisionado. Trabalho Academico.

ABSTRACT

This academic work presents an overview of the KNN algorithm (k-nearest neighbors) in the context of supervised learning. KNN is one of the simplest and most popular methods for classification and regression and is widely used in many applications. The algorithm is based on the principle that similar objects tend to be close to each other in the feature space. KNN classifies or predicts the output of an unknown object based on the classes or values of the closest training samples. In this report, we cover key concepts and a brief experiment around two databases showing execution steps and practical considerations when use the KNN and later compare with the results submitted to the WEKA using the same bases. We also discuss its advantages and limitations, as well as some strategies to improve its performance. Through this analysis, the aim is to provide a clear understanding of KNN and its key role in the field of supervised learning.

Keywords: LaTeX. UFTeX. Discipline Report. KNN. Supervised Learning Algorithm. Academic Work.

SUMÁRIO

1 INTRODUÇÃO 5

1.1 Base de Dados 5

1.2 Métodos 5

1.3 Estrutura do Experimento 6

2 O ALGORITMO KNN 7

2.1 Funcionamento 7

2.2 Sobre Iris e Wine 8

3 RESULTADOS 10

3.1 IRIS 10

3.1.1 Considerações 11

3.2 WINE 12

3.2.1 Considerações 12

1 INTRODUÇÃO

No campo do aprendizado de máquina, os algoritmos de classificação desempenham um papel fundamental na tarefa de atribuir rótulos ou categorias a objetos desconhecidos com base em um conjunto de exemplos conhecidos.

Esses algoritmos utilizam a aprendizagem supervisionada, um ramo do aprendizado de máquina, em que os dados de treinamento são compostos por pares de entrada e saída esperada.

O objetivo é encontrar um modelo capaz de generalizar a relação entre os dados de entrada e as saídas correspondentes, permitindo a classificação precisa de novos objetos.

Dentre os diversos algoritmos de aprendizado supervisionado, destaca-se o KNN (k-nearest neighbors), um método simples e intuitivo que tem sido amplamente utilizado em uma variedade de aplicações.

1.1 Base de Dados

As bases de dados usadas neste trabalho foram:

- Iris Disponível **aqui**
- Wine Disponível**aqui**

Com isso vamos considerar o número de vizinhos que o algoritmo deve considerar e uma porcentagem da base de dados a utilizar como base de conhecimento, visto que se trata de uma aprendizagem supervisionada.**Observação:** É válido ressaltar que essa base de conhecimento será construída por meio da coleta dos dados equivalente a porcentagem de forma aleatória.

1.2 Métodos

Para realizar a análise do comportamento do algoritmo e dos resultados vamos levar alguns fatores para inferir conclusões. Portanto leva-se em conta:

- Tempo de execução.
- Número de acertos.
- Taxa de precisão.

1.3 Estrutura do Experimento

Visto que ja se tem conhecimento sobre as bases de dados e os atributos a observar, em seguida usando o algoritmo KNN implementado pelos autores disponível **aqui** vamos executar usando diferentes números de vizinhos porem com taxas de base de conhecimento fixas de 30%, 50% e 70% coletar os resultados determinar padrões.

2 O ALGORITMO KNN

O KNN é baseado no princípio de que objetos semelhantes tendem a estar próximos uns dos outros no espaço de características. Assim, ele classifica ou prevê a saída de um objeto desconhecido com base nas classes ou valores das amostras de treinamento mais próximas a ele.

Compreender o KNN é fundamental para a compreensão geral dos algoritmos de aprendizado supervisionado, uma vez que suas características básicas podem ser estendidas e aprimoradas em algoritmos mais complexos.

2.1 Funcionamento

O algoritmo pega uma parte de um conjunto de dados rotulados e compreende como seu conhecimento que servirá de base e em seguida leva em conta a parte restante do conjunto de teste.

Para cada item no conjunto de teste o algoritmo faz uma previsão de qual será o rótulo dele com base na proximidade do elemento atual com os seus K-Vizinhos.

Quanto mais próximo ele for de seu vizinho significa que há uma maior probabilidade dele pertencer ao conjunto do mesmo.

```
1 def euclidian (testPoint, trainPoint) :
2
3     dist = 0
4
5     for i in range (len (testPoint)) :
6
7         dist = dist + (testPoint [i] - trainPoint [i]) ** 2
8
9     return dist ** 0.5
```

Essa previsão é heurística, isto é, leva em consideração uma estratégia para guiar. Essa estratégia será responsável por definir o grau de similaridade do elemento atual e dos seus K-Vizinhos mais próximos.

Esse cálculo leva em consideração os atributos necessários, e assim calcular a distância euclidiana entre eles para realizar a comparação

Figura 1 – Fórmula Distancia

$$\sqrt{\sum_{i=1}^{n}(p_i - p_q)}$$

Posteriormente uma lista dos K-Vizinhos será gerada com a finalidade de mapear os mesmo e a classe que mais aparecer será escolhida como rótulo do elemento previsto em questão.

2.2 Sobre Iris e Wine

- IRIS

  A flor de iris é uma flor roxa que tem vários tipos e convenientemente há um conjunto de dados disponível na internet que possui informações rotuladas sobre três tipos dessa flor (setosa, versicolor e virginica). O objetivo do KNN será prever qual o tipo de uma iris baseando-se nas informações sobre ela.

  A base possui 150 linhas de informações, 4 colunas de atributos e 1 coluna para o rótulo.

  É uma quantidade pequena de dados e portanto hora teremos uma pequena base de dados e uma grande base de teste e assim vice e versa.

  Logo, ou a precisão vai ser grande mas com poucos testes ou a precisão vai ser baixa com muitos testes.

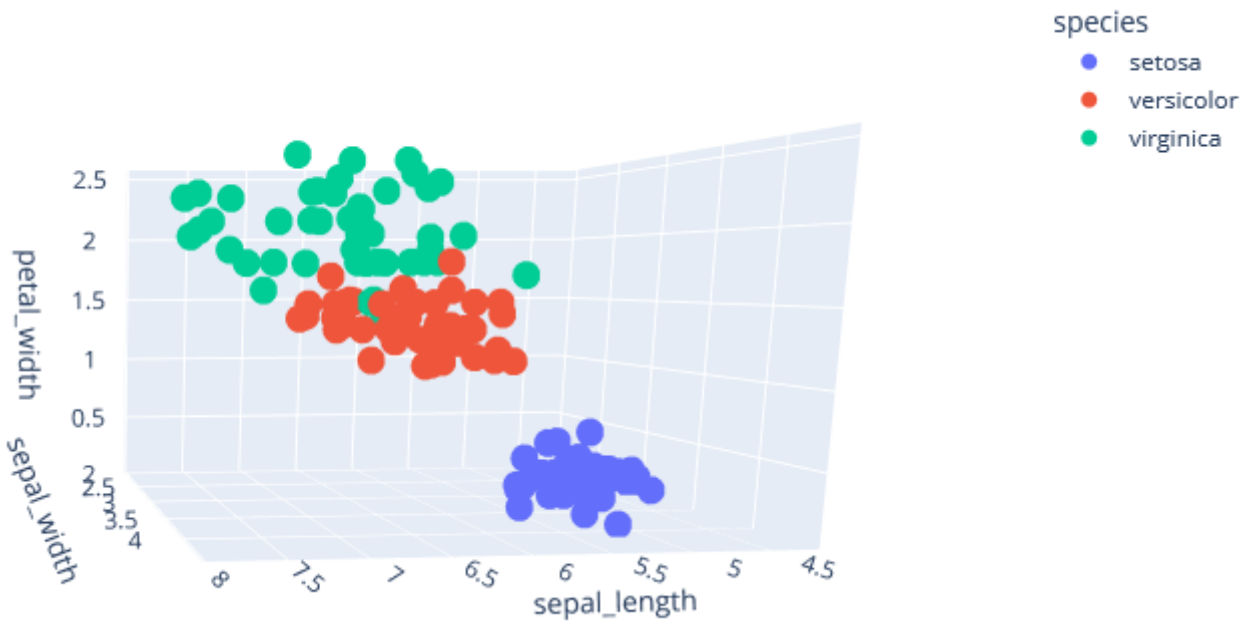

Figura 2 – Visão espacial da base de dados Iris

- WINE

  Contém 13 colunas referentes a atributos e 1 coluna como o rótulo da classe.

  Possui 178 linhas, um pouco maior que a IRIS e ainda possui o mesmo problema de ter que balancear a porcentagem dos dados que será utilizado para a base de conhecimento e para os testes.

3 RESULTADOS

Por meio desta análise, esperamos fornecer aos leitores uma visão clara e abrangente do KNN, destacando sua relevância e contribuições para o campo do aprendizado de máquina.

3.1 IRIS

Tabela 1 – IRIS - 30%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	30%	61.9047619%	65	0.03128337860
2	30%	60.95238095%	64	0.04685759544
3	30%	63.80952380%	67	0.06599974632
4	30%	63.80952380%	67	0.07816982269
5	30%	63.80952380%	67	0.09380412101
7	30%	60.95238095%	64	0.15233898162

Tabela 2 – IRIS - 50%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	50%	97.333 %	73	0.031210184097
2	50%	97.333 %	73	0.051963806152
3	50%	98.66%	74	0.078118085861
4	50%	98.66%	74	0.097542762756
5	50%	98.66%	74	0.125
7	50%	98.66%	74	0.147415161132

Tabela 3 – IRIS - 70%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	70%	69.56521739 %	32	0.03000116348
2	70%	69.56521739 %	32	0.05254244804
3	70%	69.56521739%	32	0.06249499320
4	70%	71.73913043%	33	0.07808256149
5	70%	71.73913043%	33	0.10681867599
7	70%	71.73913043%	33	0.11638212203

3.1.1 Considerações

O algoritmo se sai muito bem em todos os casos tendo seu melhor desempenho quando tem conhecimento de metade da base de dados (50% no caso). E percebemos que o número de acertos quando passamos 70% como base de conhecimento é relativamente baixo apesar da alta taxa de precisão pois a maior parte dos dados vai pertencer ao conjunto de conhecimento, restando apenas poucas tuplas para o conjuntos de testes.

3.2 WINE

Tabela 4 – WINE - 30%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	30%	68%	85	0.07551312446
2	30%	68.8%	86	0.12505483627
3	30%	66.4%	83	0.19931650161
4	30%	64.8%	81	0.25952792167
5	30%	68%	85	0.31325244903
7	30%	64.8%	81	0.44710636138

Tabela 5 – WINE - 50%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	50%	96.62921348%	86	0.07551312446
2	50%	92.13483146%	82	0.12505483627
3	50%	93.25842696%	83	0.19931650161
4	50%	87.64044943%	78	0.25952792167
5	50%	93.25842696%	83	0.31325244903
7	50%	92.13483146%	82	0.44710636138

Tabela 6 – WINE - 70%.

K	% de Base de conhecimento	Precisão	N° de acertos	Tempo de execução (s)
1	70%	81.48148148%	44	0.07551312446
2	70%	68.51851851%	37	0.12505483627
3	70%	74.07407407%	40	0.19931650161
4	70%	66.66%	36	0.25952792167
5	70%	70.37037037%	38	0.31325244903
7	70%	62.96296296%	34	0.44710636138

3.2.1 Considerações

O algoritmo atinge uma precisão estável e de certa forma alta dada a porcentagem de 30% como base de conhecimento, ou seja, com poucos dados em sua base foi capaz de realizar boas previsões, atingiu um comportamento relativamente ótimo quando usado 50% como base de conhecimento, porém quando se tem a maior parte do conjunto de dados em sua base de conhecimento ele assume umas postura instável variando muito seus resultados pois a base de dados é consideravelmente maior, então... por consequência as vizinhanças são maiores entre si.