Comparação de Técnicas de Aprendizado de Máquina

Felipe Samuel, Gabriel Carneiro, Cristian Ribeiro

Resumo

Com o avanço da tecnologia, novas técnicas e métodos vieram sendo desenvolvidos no decorrer dos anos a fim de solucionar o grande volume de dados gerados gradativamente com o crescimento populacional e o acesso a internet. Portanto, é neste contexto que esta pesquisa atua, analisando as diferentes aplicações de algoritmos de aprendizado supervisionado em diferentes bases de dados de modo a comparar os resultados obtidos através de uma validação cruzada, dito isso pode-se perceber que os algoritmos não mantiveram excelência em todos as induções, isto é, não existe um algoritmo ideal para todos os caso, é preciso analisa-los separadamente e especificamente para determinada base.

Palabras clave: aprendizado supervisionado. base de dados. validação cruzada.

[1] Faculdade Federal de São João del Rei, Campus Tancredo de Almeida Neves, Brasil.
[2] Sistemas Informáticos de Inteligencia Artificial, Professor Edmílson Batista dos Santos.

1. Introdução

Um sistema de Inteligência Artificial não é capaz somente de armazenamento e manipulação de dados, mas também da aquisição, representação, e manipulação de conhecimento (GINAPE, 2021). Dito isto o propósito deste trabalho é e aplicar diferentes algoritmos de aprendizado supervisionado em distintas bases de dados com o objetivo de comparar seus resultados utilizando um método de validação cruzada.

1.1. Preparação

Para gerar a indução nos algoritmos de classificação as bases de dados foram escolhidas de acordo com a significância e quantidade de acesso obtida no repositório UCI, serão dispostas da seguinte forma: Os algoritmos escolhidos para análise da base de dados previamente especificadas, para a indução será:

- Árvore de Decisão
- Vizinhos mais Próximos (K-Nearest Neighbor – KNN)
- Naive-Bayes
- Redes Neurais
- Máquina de Vetor de Suporte (Support Vector Machines – SVM).

2. Algoritmos de Classificação

Para atingirmos um melhor entendimento se mostra necessário entender o funcionamento de cada classificador e quais parâmetros estão sendo variados para encontrar a melhor acurácia.

2.1. Arvore de Decisão

Árvores de Decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados (Maia, Gomes, Chagas, 2017). Basicamente um algoritmo de classificação inspirado em árvore de decisão, utiliza o método dividir para conquistar na hora de resolver um problema de decisão. Para encontramos a melhor acurácia fizemos uma variação no tamanho mínimo das folhas, o que gera uma variância maior nos resultados.

2.2. Vizinhos mais Próximos (K-Nearest Neighbor – KNN)

O algoritmo K-Nearest Neighbor, possui aprendizagem baseada em instâncias, os algoritmos desta família armazenam todas as instâncias de treinamento e quando uma nova instância é apresentada ao algoritmo para ser classificada, um conjunto de instâncias similares a essa nova instância é recuperada do conjunto de treinamento e utilizada para classificá-la (FARIA,2016). Para encontramos a melhor acurácia fizemos uma variação no número de vizinhos próximos.

2.3. Naive-Bayes

O classificador Naïve Bayes é baseado na suposição simplificadora de que os valores dos atributos são condicionalmente independentes dado o valor alvo. Ou seja, a probabilidade de observar a conjunção de atributosa1,a2,...,an é somente o produto das probabilidades para os atributos individuais.(KOERICH, 2021). Para encontra a melhor acurácia fizemos uma variação no atributo ẗvar_smoothing; que gera uma grande variação nos resultados e uma probabilidade maior de ter uma acurácia alta.

2.4. Redes Neurais

Redes neurais buscam implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas. Nesse sentido, apresentam capacidade de adaptar os seus parâmetros como resultado da interação com o meio externo, melhorando gradativamente o seu desempenho na solução de um determinado problema(FERNEDA, 2016). Para encontrar a melhorar acurácia variamos o número de hidden layers(camadas de nerônios), em algumas bases de dados foi necessário que a variação fosse menor por causa da especificidade dos dados.

2.5. Máquina de Vetor de Suporte (Support Vector Machines – SVM).

Têm a capacidade de resolver problemas de classificação e regressão, adquirindo com o aprendizado na etapa de treinamento a capacidade de generalização[...] O objetivo é produzir um classificador que funcione de forma adequada com exemplos não conhecidos[...] Adquirindo assim a capacidade de predizer as saídas de futuras novas entradas.(JÚNIOR, PRUDÊNCIO, 2010). Para obtermos uma melhor acurácia variamos o parâmetro C (Regularization parameter), a intensidade da regularização é inversamente proporcional a C que deve ser estritamente positiva.

3. Banco de dados

3.1. Iris

O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de íris. Uma classe é linearmente separável das outras 2; os últimos NÃO são linearmente separáveis uns dos outros.

```
Características dos dados:  Multivariada
Número de Instâncias:178
Área:Física
Características do atributo: Inteiro, Real
Número de atributos: 13
Data Doada: 01-07-1991
Tarefas associadas: Classificação
Valores ausentes: Não
Número de acessos na web: 1642176

Informações de atributo:
1. comprimento da sépala em cm
2. largura da sépala em cm
3. comprimento da pétala em cm
4. largura da pétala em cm
5. classe:
- Iris Setosa
- Iris Versicolour
- Iris Virginica
```

3.2. Dígitos

Os bitmaps de 32x32 são divididos em blocos não sobrepostos de 4x4 e o número de pixels on é contado em cada bloco. Isso gera uma matriz de entrada de 8x8, onde cada elemento é um número inteiro no intervalo de 0 a 16. Isso reduz a dimensionalidade e dá invariância a pequenas distorções.

```
Características dos dados: Multivariada
Número de Instâncias: 5620
Área: Computador
Características do atributo: Inteiro
Número de atributos: 64
Data Doada: 01/07/1998
Tarefas associadas: Classificação
Valores ausentes: Não
Número de acessos na web: 303326

Informações de atributo:
Todos os atributos de entrada são inteiros
no intervalo de 0 a 16.O último atributo é
o código de classe 0..9.
```

3.3. Vinho

Esses dados são resultados de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de três cultivares diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos.

```
Características dos dados: Multivariada
Número de Instâncias: 178
Área: Física
Características do atributo: Inteiro, Real
Número de atributos: 13
Data Doada: 01-07-1991
Tarefas associadas: Classificação
Valores ausentes: Não
Número de acessos na web: 1642176

Os atributos são:
1) Álcool
2) Ácido málico
3) Cinzas
4) Alcalinidade das cinzas
5) Magnésio
6) Fenóis totais
7) Flavonoides
8) Fenóis não flavonoides
9) Pro antocianinas
10) Intensidade de cor
11) Matiz
12) OD280 / OD315 de vinhos diluídos
13) Prolina
```

3.4. Câncer de mama

Os recursos são calculados a partir de uma imagem digitalizada de um aspirado por agulha fina (FNA) de uma massa mamária. Eles descrevem as características dos núcleos celulares presentes na imagem.

```
Características dos dados:  Multivariada
Número de Instâncias: 569
Área: Vida
Características do atributo: Real
Número de atributos: 32
Data Doada: 01/11/1995
Tarefas associadas: Classificação
Valores ausentes: Não
Número de acessos na web: 1474312

Dez características de valor real são calculadas para
a) raio (média das distâncias)
b) textura (desvio padrão)
c) perímetro
d) área
e) suavidade (variação local)
f) compacidade (perímetro ^ 2 / área - 1,0)
g) concavidade (severidade das porções)
h) pontos côncavos (número de porções)
i) simetria
j) dimensão fractal ("aproximação do litoral" - 1)
```

4. Análise de dados

Para facilitar o entendimento dos gráficos a seguir, tomaremos as seguintes definições:
Accuracy: Porcentagem de acertos em determinados parâmetros.
N: Variação de parâmetros específicos para cada algoritmo, explicados previamente na seção 2.

4.1. Análise dos algoritmos de classificação sobre a base de dados: Íris

Durante o supervisionamento dos algoritmos de aprendizado na base de dados Íris, percebemos que o conjunto compartilha de uma acurácia máxima semelhante, todos tiveram bons resultados se aproximando da taxa de acerto ideal de 100%, porém, o algoritmo SVM, se mostra mais eficaz e mais constante, quando o parâmetro C (Regularization parameter) está entre 0 e 50. A intensidade da regularização é inversamente proporcional a C, sendo estritamente positiva conforme a **Figura 1**.

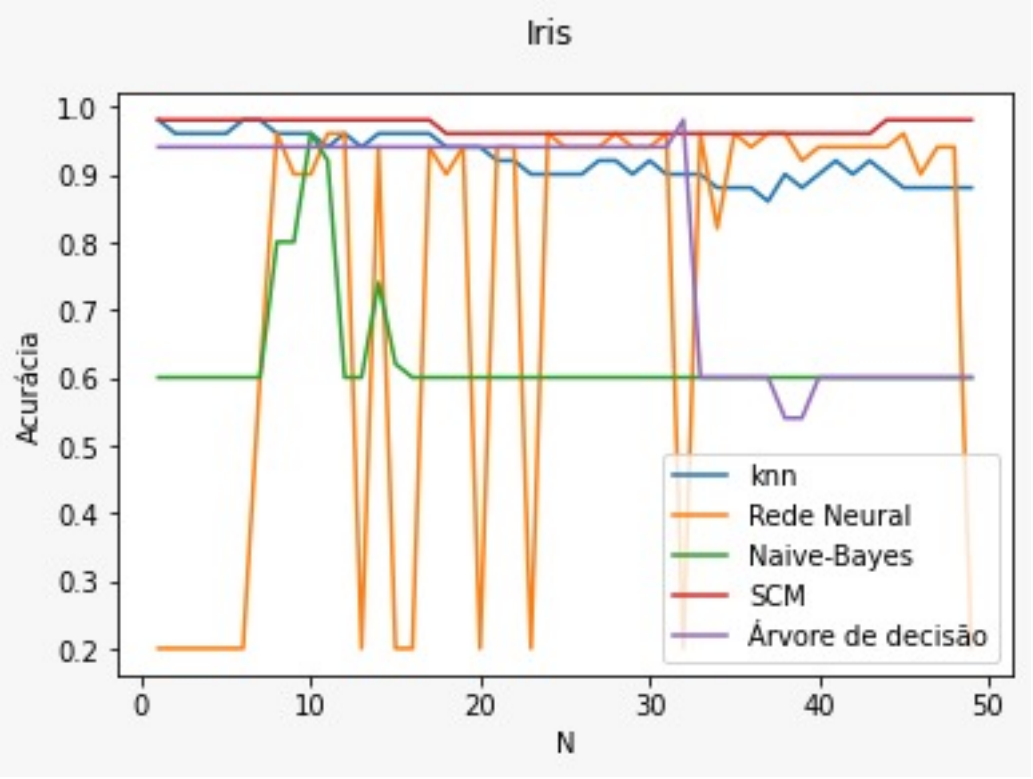

Figura 1. Indoção dos algoritimos de classificação sobre a base: Íris

Vale dizer que a Máquina de vetor de suporte, sustenta a acurácia no seu estado máximo tanto para parâmetros próximos de 0, quanto para parâmetros próximos de 50, manteve uma estabilidade incrível, se mantendo sempre acima da taxa de acerto de 90%, Todavia, o algorítimo que atingiu a menor acurácia foi a Rede Neural, oscilando seus picos de máximos e mínimos(acurácia) durante toda variação dos parâmetros.

4.2. Análise dos algoritmos de classificação sobre a base de dados: Dígitos

Ao observamos as análises dos dados da base Dígitos, percebemos uma similaridade interessante entre o algorítimo SCM e KNN durante a indução, ambos tiveram uma leve queda (acurácia), na medida que os parâmetros iam se afastando de zero, Entretanto vale dizer que os algoritmos obtiveram resultados satisfatórios, acima de 80%. Dito isto, mesmo obtendo valores similares ao KNN, o SCM foi o que apresentou o melhor desempenho partindo de 99% de acurácia, dada a variação obtida através de n(número de vizinhos próximos).

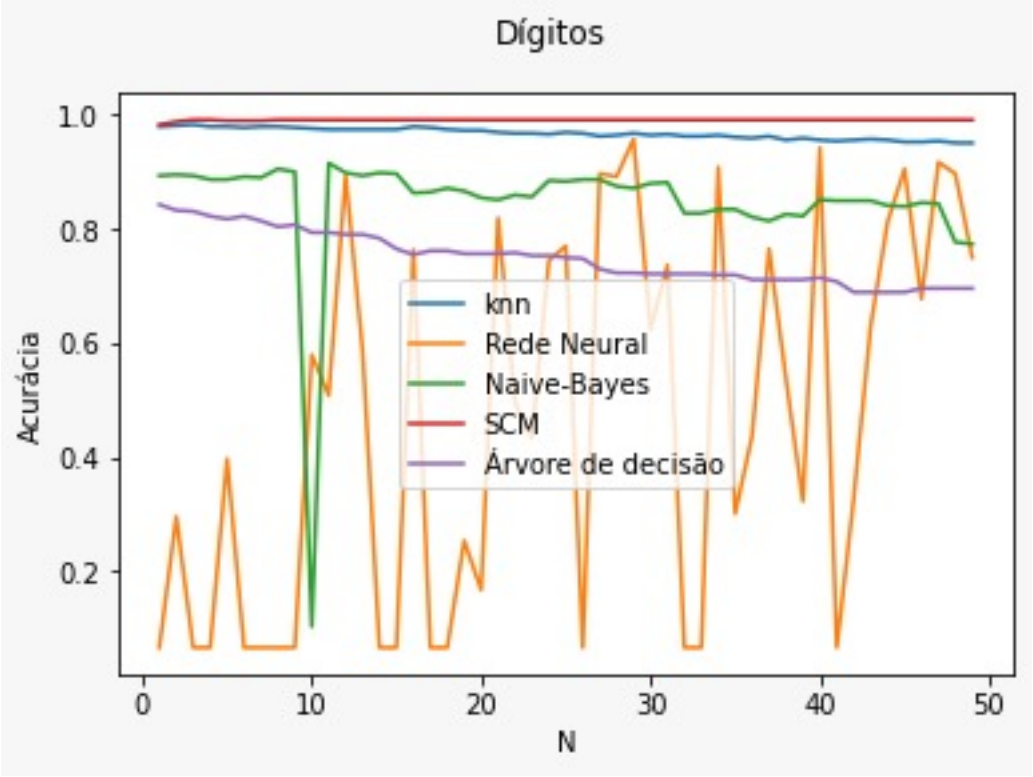

Figura 2. Indução dos algoritimos de classificação sobre a base: Digitos

E mantendo o nível de excelência desde parâmetros próximos a 0 até parâmetros próximos a 50, é possível perceber, que alguns algorítimos de classificação se comportam da mesma forma, com exceção da Rede neural, começam com uma acurácia alta e vão perdendo precisão a medida que os parâmetros variam

4.3. Análise dos algoritmos de classificação sobre a base de dados: Vinho

Através dos dados obtidos na análise dos algoritmos da base Vinho, é possível perceber que a taxa de acertos da Rede Neural foi insustentável, cerca de 40% no melhor resultado, é possível dizer também que KNN e Scm não se comportaram de forma semelhante, porém com uma taxa de acertos inferior a 90% e 80%, Naive-Bayes se mostrou mais efetivo durante a indução desta base, tendo seu máximo quando o parâmetro atingiu 10, dada a variação no atributo "var_smoothing".

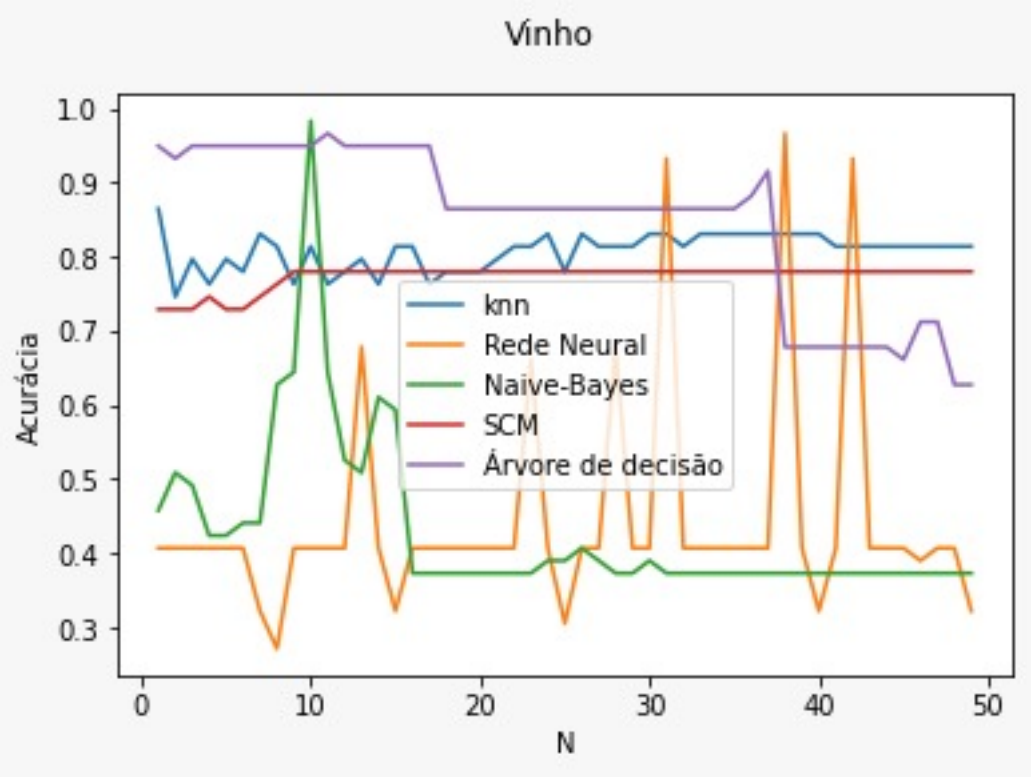

Figura 3. Indoção dos algoritimos de classificação sobre a base: Vinho

Outra análise a ser feita, foi a excelência no comportamento da Árvore de decisão, tendo uma acurácia significativa, para parâmetros entre 0 e 20, relativo a variação no tamanho mínimo das folhas, obtendo uma variância maior nos resultados, com aproximadamente 95% de acurácia cada.

4.4. Análise dos algoritmos de classificação sobre a base de dados: Câncer de Mama

Durante o supervisionamento dos algoritmos de aprendizado na base de dados Câncer de Mama, percebemos comportamento semelhante entre 3 algoritmos: Knn, SCM, Arvore de decição, vale dizer também que todos algoritmos, tiveram bons resultados acima de 90% de acerto dada variação dos parâmetros.

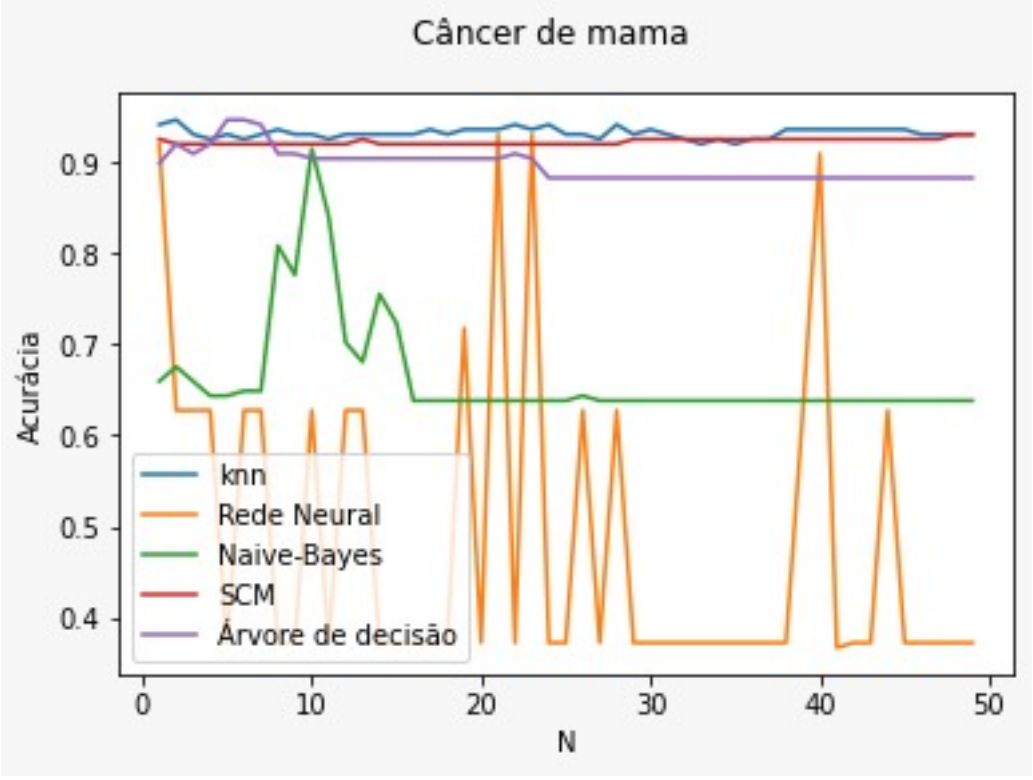

Figura 4. Indoção dos algoritimos de classificação sobre a base: Câncer de Mama

4.4.1. Análise geral

Após analisar os diferentes gráficos, percebemos uma similaridade no comportamento de alguns algoritmos. A Rede Neural, manteve-se oscilando em todos as bases de dados, com picos de máximos e mínimos consequente a variação dos parâmetros, também foi a única que atingia valores mínimos para parâmetros iguais a zero, pode-se perceber que o tempo de execução de cada algoritmo oscilou dependendo da base de dados, sendo que a base dígitos consumiu um maior tempo.

5. Conclusão

O objetivo deste trabalho foi desenvolver e aplicar diferentes algoritmos de aprendizado supervisionado às diferentes bases de dados de modo a comparar seus resultados. É plausível dizer que no decorrer do trabalho entender como fazer a base de dados operar nos algoritmos específicos não foi o maior desafio, o obstáculo encontrado foi a implementação do Vizinhos mais Próximos (K-Nearest Neighbor – KNN). Dito isto, durante o desenvolvimento, conseguimos atingir o objetivo e portanto, este trabalho abre uma uma oportunidade de simular e compreender a indução de vários algoritmos de classificação em diversas bases de dados, vale dizer que nenhum algoritmo foi melhor em todos os casos, é necessário um estudo para determinar qual seria o mais indicado.

6. Referências

1. Wolberg, Street, Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set" Machine Learning Repository's, 1993, 1994, 1995. [Online]. Disponível: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisc

onsin+(Diagnostic)

2. E. Alpaydin, C. Kaynak, "Optical Recognition of Handwritten Digits Data Set" Machine Learning Repository's, 1998, 1995. [Online]. Disponível: https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

3. R.A. Fisher, Michael Marshall, "Iris Data Set" Machine Learning Repository's, 1993, 1950, 1972, 1973, 1980, 1988. [Online]. Disponível: https://archive.ics.uci.edu/ml/datasets/Iris

4. Forina, M. et al, PARVUS, "Wine Data Set" Machine Learning Repository's, 1992. [Online]. Disponível: https://archive.ics.uci.edu/ml/datasets/Wine

5. Latex Editor, "Overleaf" 2021. [Online]. Disponível: https://pt.overleaf.com/learn

6. Python, "python library", supervised-learning, 2007,2020. [Online]. Disponível: https://scikit-learn.org/