Avaliação do Desempenho das Técnicas de Mineração de Dados num Dataset de Diagnósticos de Doenças Eritemato Escamosas

Matheus Oliveira
Sistemas de Informação
Instituto Federal do Espírito Santo (IFES)
Serra, Brasil
matheussouzapoliveira@gmail.com

***Resumo*—Este artigo tem o objetivo de apresentar os resultados obtidos utilizando técnicas de Mineração de Dados (*Data Mining*) e com um pouco de *Machine Learn*, sobre um dataset de doenças Eritemato Escamosas, doenças de pele que possuem sintomas semelhantes, e definir qual o melhor para esta situação. Dois algoritmos foram utilizados um classificador e um agrupador o algoritmo de classificação escolhido foi a Árvore de Decisão e o algoritmo de agrupamento o K-médias, e com o resultado foi possivel concluir que o algoritmo de Árvore de Decisão se saiu melhor com uma acuráciad 97,96%**

***Palavras-Chave*—Algoritmos, Doenças de Pele, Machine Learn, Scikit Learn, Classificação, Agrupamento**

I. Introdução

Com o aumento expressivo da quantidade de dados amazenados por empresas e organizações, seus profissionais passaram a se preocupar em como extrair informações das imensas bases de dados. Nos anos 80 inicia-se um novo conceito chamado de *Data Mining* (DM ou Mineiração de Dados), que tem como principal objetivo o tratamento dessas bases de dados, buscando alguma relação seus elementos, a fim de se obter alguma tendência futura ou padrões de comportamento. Dito isto, é possivel perceber a importância desse ramo da tecnologia para as empresas [1].

Data Mining significa em seu sentido literal minerar os dados em busca de extrair algo de valor deles, a mineração de dados é uma etapa do *Knowledge Discovery in Databases*, ou simplismente KDD. O KDD é um processo que consiste em algumas etapas, são elas: 1. limpeza dos dados, 2. integração dos dados, 3. seleção, 4. transformação dos dados, 5. **Mineração**, 6. avalialção e 7. vizualização dos resultados. Seu objetivo, como o da mineiração de dados, é justamente conseguir descobrir algum conhecimento em um Banco de Dados [1].

Existem algumas técnicas para a realização da mineração de dados, dentre elas está a, classificação supervisiodada, que se trata de um procedimento onde se determina um mapeamento capaz de indicar a qual classe pertence um objeto, a partir de um conjunto de dados previamente classificados, as principais metodologias de classificação são: árvore de decisão, regressão linear e regressão logistica. Outra técnica de DM é o agrupamento (ou *clustering*), que diferentemente da classificação este método divide uma base de dados em grupos que possuem características semelhantes entre si, nesse caso não é necessário um conjunto antes classificado, exemplos de métodos de agrupamentos, agrupamento em árvore e agrupamento por k-Médias, o famigerado *k-Means* [2]. A fim de avaliar um método de classificação e outro de agrupamento, foram escolhidos os métodos de Árvore de Decisão, para classificação e o *k-Means*, para avaliação do agrupamento.

O dataset análisado neste trabalho se refere à uma base sobre doenças eritemato escamosas, doenças de pele que são um problema para a dermatologia, pois apresentam sintomas e características similares em seus estágios iniciais e para um mesmo paciênte é possivel obter diferentes diagnosticos entre médicos distintos.

O Python foi utilizado para a aplicação de todas etapas necessárias para a avaliação das técnicas, juntamente à biblioteca *Scikit-learn* onde estão os algoritmos que representam as metodologias citadas anteriormente, o Pandas foi utilizado para o pré-processamento e manipulção dos dados, o matplotlib para a plotagem dos gráficos, utilizados para a análise exploratória e análise dos resultados obtidos, dentre outras bibliotécas.

II. Referencial Teórico

A mineração de dados auxilia na descoberta de padrões de comportamento entre as entidades das gigantes bases de dados, suas técnicas permitem a avaliação dessas bases de forma prática. No caso deste trabalho busca-se utilizar métodos de classificação e agrupamento para avaliar pessoas com diferentes doenças de pele, mas que possuem sintomas semelhantes.

A classificação é uma técnica preditiva que consiste na análise de dados em busca da definição de padrões. Essa atividade define um conjunto de modelos que podem ser utilizados para classificar novos objetos, ou seja, para a criação de um modelo precisamos antes de um conjunto de dados de amostragem, corretamente classificados. No caso deste trabalho iremos utilizar a estratégia de árvore de decisão, ela é capaz de representar o conhecimento em forma de árvore, este conhecimento pode ser visto também na forma de regras de classificação do tipo SE-ENTÃO, onde cada nó

filho representa uma condição envolvendo um atributo e um conjunto de valores possíveis, no caso das folhas da árvore elas representam a conclusão daquele ramo da árvore [5].

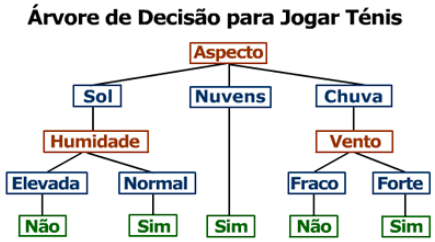

Figura 1. Exemplo de uma árvore de decisão simples, que decide se o dia está bom para jogar Ténis.

Agrupamentos são grupos formados por objetos que compartilhem características comuns. Agrupar significa dividir os elementos em grupos. Posto isto, agrupamento (ou *clustering*) é outra técnica de mineração de dados para realizar agrupamentos automáticos, o agrupamento é considerado uma técnica de classificação não supervisionada, porque não necessita-se de classes pré-definidas, contrariando o método de classificação supervisionada citado no parágrafo anterior [6].

O método de agrupamento aplicado neste trabalho é conhecido como agrupamento particional, em que a divisão dos objetos do conjunto em subconjuntos não interligados de como que cada objeto esteja em um único subconjunto. O K-médias (*K-means*) fornece uma classificação de informações de acordo com os próprios dados, ele funciona da seguinte forma, são definidos centroides, nomalmente a partir da média dos padrões do grupo, o número de K de centroides é escolhido no inicio pelo usuário, definido a quantidade de grupos cada objeto é atribuido à centroide mais próxima [6].

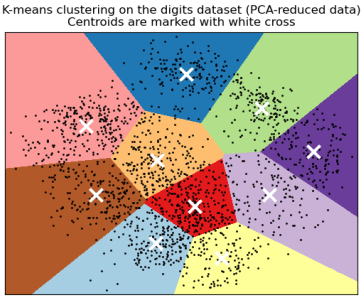

Figura 2. Exemplo de um agrupamento utilizando o K-médias.

O Scikit-learn é uma bibliotéca *open-source* para machine learn em Python, segundo sua própria documentação, é uma bibliotéca simples e eficiente com ferramentas para predição e análise de dados, baseada em NumPy, SciPy e matplotlib, outras poderosas bibliotécas de Python também utilizadas neste trabalho. Através do sklearn temos os algoritmos de árvore de decisão e o k-médias [7].

Outra bibliotéca *open-source* famosa do Python utlizada aqui foi o pandas, que é uma biliotéca para análise e manipulação de dados de alta performance, com ela foi possivel obter todo suporte para o tratamento e manipulação dos dados utilizados em nossa análise [8].

III. Metodologia

Após a obtenção do dataset foi criado o arquivo *columns.json* que representava as colunas da base de forma mais descritiva, para um melhor entendimento dos dados, este arquivo foi criado com base na descrição do arquivo de dados, que também disponibilizado anteriormente e todos foram enviados ao Github. O dataset analisado descreve uma série de pessoas com seis doenças distintas: *Psoriasis* (1), *Seboreic Dermatitis* (2), *Lichen Planus* (3), *Pityriasis Rosea* (4), *Cronic Dermatitis* (5), *Pityriasis Rubra Pilaris* (6).

Além disso, os demais atributos da base descreviam sintomas e poderiam possuir valores inteiros de 0 à 3 de acordo com sua intensidade, exceto a coluna *age*, idade do paciênte, e *family_history*, histórico familiar, onde 1 indica que possuia histório na família e 0 que não,

Todo o ambiente foi preparado através do Google Colaboratory, o Colab é uma forma de executar notebooks em Python pelo navegador. Primeiramente importamos todas as bibliotécas necessárias incluindo o pandas e o sklearn, em seguida foram lidos os arquivos de colunas e a base que continha os dados das doenças eritemato escamosas, após toda preparação o pré processamento dos dados foi iniciado.

De acordo com a descrição do dataset haviam três linhas que não possuiam idades e os valores dessas idades foram substituidos pelo caractere '?', para remover essa inconsistência nos dados esse caracter foi substituido pela média das idades do dataset.

Uma análise de *outliers* em relação as idades dos paciêntes foi feita, por um gráfico tipo box-plot, mostrado pela Figura 3, e por ela foi possivel observar que os dados estavam bem distribuidos, somente um *outlier* foi encontrado para a doença 6.

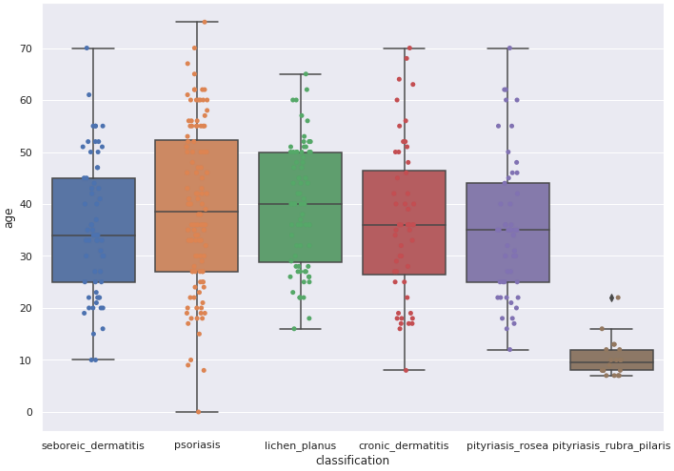

Figura 3. Gráfico box-plot criado a partir dos dados utilizados.

Quanto ao balanceamento dos dados foi criado um gráfico de barras (Figura 4) para analisar a distribuição dos paciêntes pelas doenças, por esse gráfico foi possivel observar um grande discrepância, para menos, na quantidade de dados para a doença 6 com somente vinte ocorrências e para mais, na doença 1 com 112 instâncias. Portanto, foi decidido manter para cada doença 61 ocorrências (média das ocorrências), então realizamos undersampling nas doenças com mais de 61 ocorrências e oversampling nos dados com menos.

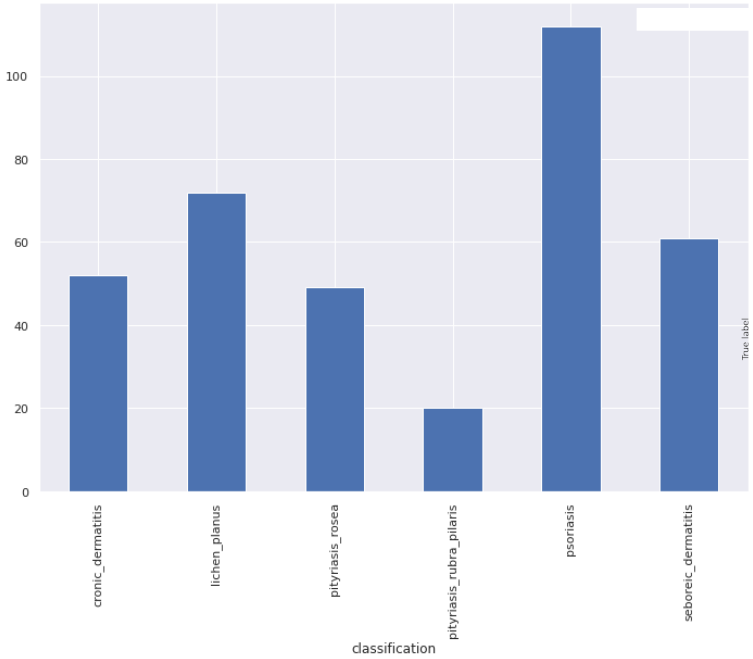

Figura 4. Gráfico de barras da quantidade de doenças x quantidade de ocorrências.

Finalizado o pré processamento criamos um classificador e o agrupador utilizando os algoritmos de árvore de decisão (*DecisionTreeClassifier*) e o de k-médias (*KMeans*) respectivamente, obtidos do Scikit Learn. Para o *DecisionTreeClassifier* separamos 60% dos dados para treino e 40% dos dados para testes. Para *KMeans* excluirmos a coluna *classification*, pois devemos considerar dados não classificados e definimos que deveriam ter 6 grupos distintos já que anteriormente tinhamos 6 doenças.

IV. Resultados

A. Árvore de Decisão

Para o *DecisionTreeClassifier* obtivemos um ótimo resultado de acurácia, tanto para o grupo de treino, quanto para o grupo de teste, foi possivel obter uma acurácia de 100% para a base de treino, e 97,96% no teste.

Com a base de teste foi criada a matriz de confusão, apresentada na Figura 5, e com o K-folds dividi-lá em cinco partes para avaliar a acurácia média do modelo nessas diferentes partes e a acurácia média do nosso classificador foi de 93,26%, um resultado um pouco menor do que quando a base de teste era somente uma.

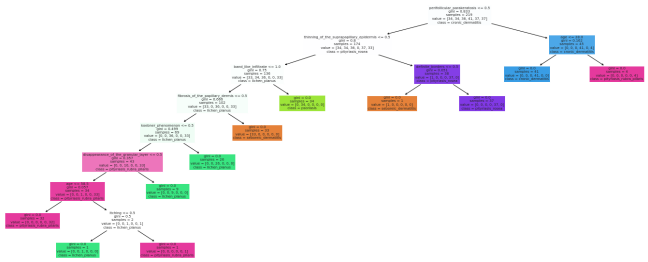

Figura 5. Árvore de decisão obtida a partir do *DecisionTreeClassifier*.

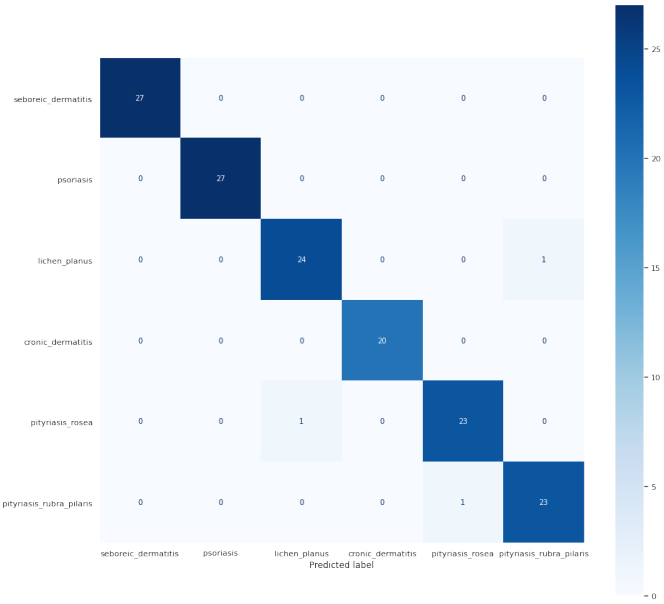

Figura 6. Matriz de confusão.

B. K-médias

Ao contrário do algoritmo anterior o *K-means* não foi tão bem, comparando os dois. O algoritmo resultou em seis grupos, definidos antes de sua execução, número ecolhido devido à quantidade de doenças analisadas no trabalho.

Para avalia-lo vamos considerar os grupos do K-means e os originais, ordenados em ordem decrescente de ocorrências se fizermos uma relação direta, ou seja, o grupo com maior número de ocorrencias no K-means é a doença com maior número de ocorrências na base disponibilizada. Considerando isso como verdadeiro tivemos o seguinte resultado, mostrado na Figura 7.

Grupo K-means	Ocorrências	Grupos Originais	Ocorrências	Δ
1	88	Psoriasis	112	-21%
2	77	Lichen Planus	72	7%
3	64	Seboreic Dermatitis	61	5%
4	56	Cronic Dermatitis	52	8%
5	53	Pityriasis Rosea	49	8%
6	28	Pityriasis Rubra Pilaris	20	40%

Figura 7. Comparação grupo K-means com os grupos originais.

V. Conclusão

É claro pela análise dos resultados que algoritmo de árvore de decisão teve um resultado melhor do que o K-means, enquanto o agrupador classificou 318, elementos de forma correta nos dando uma taxa de acerto de 86,89%, levando em conta a Figura 7, o classificador com uma acurácia de 97,96% foi superior. Entretanto, isso não quer dizer que um algoritmo é melhor que o outro, mas nesse caso a escolha para analisar proximos casos de doenças como essa seria o de árvore de decisão devido à sua maior precisão

Com este trabalho conseguimos observar o poder das tecnologias disponibilizadas atualmente, quando utilizadas de forma correta. Um trabalho que antes era feito por médicos hoje pode ser feito por uma máquina, é claro, dando um tratamento de forma certa. Utilizando técnicas de *Data Mining* e *Machine Learn*, e possivel analisarmos grandes bases de dados para prevemos padrões e tendências futuras e isso é algo de grande valor tanto para o mercado quanto para a sociedade.

Referências

[1] Sandra Amo, "Técnicas de Mineração de Dados", Universidade Federal de Uberlândia.

[2] Alex L. Ramos, Cícero N. dos Santos, "Combinando Algoritmos de Classificação para Detecção de Intrusão em Redes de Computadores", Universidade Federal de Minas Gerais, Março-2009.

[3] Rafael Santos, "Introdução à Mineração de Dados com Aplicações em Ciências Ambientais e Espaciais", Instituto Nacional de Pesquisas Espaciais.

[4] Prof. Dr. rer.nat. Aldo von Wangenheim, Análise de Agrupamentos, Disponível em: https://www.inf.ufsc.br/ aldo.vw/patrec/agrupamentos.html, Acesso em: 25/08/2021.

[5] Diana Colombo Pelegrin, Diego Paz Casagrande, Merisandra Côrtes de Mattos, Priscyla Waleska Targino de Azevedo Simões, Rafael Charnovscki, Jane Bettiol, "As Tarefas de Associação e de Classificação na Shell de Data Mining Orion", Universidade do Extremo Sul Catarinense.

[6] Jakelson Carreiro Mendes, "Agrupamento de dados e suas aplicações, Universidade Federal do Maranhão", Julho-2017.

[7] Scikit Learn, "About us", Disponivel em: https://scikit-learn.org/stable/about.html, Acesso em: 29/08/2021.

[8] pandas, "About us", Disponivel em: https://pandas.pydata.org/about, Acesso em: 29/08/2021.