VI Encontro de Iniciação Científica e Tecnológica
VI EnICT
ISSN: 2526-6772
IFSP – Câmpus Araraquara
21 e 22 de outubro de 2021

Aplicação de algoritmos de classificação para apoio à tarefa de mineração de dados

GABRIEL LEONI DUARTE[1], MARCELO CRISCUOLO[2], FÁBIO JOSÉ JUSTO DOS SANTOS[3], TALES BOALIM[4]

[1] Discente do Técnico em Informática Integrado ao Ensino Médio, Bolsista PIBIFSP, IFSP Câmpus Araraquara, gabriel.leoni@aluno.ifsp.edu.br
[2 3] Docente do Instituto Federal de São Paulo - Câmpus Araraquara, criscuolo@ifsp.edu.br, fabiojjs@ifsp.edu.br
[4] Gerente de Inovação - JN Moura Informática, tales@jnmoura.com.br

Área de conhecimento (Tabela CNPq): Banco de Dados – 1.03.03.03-0

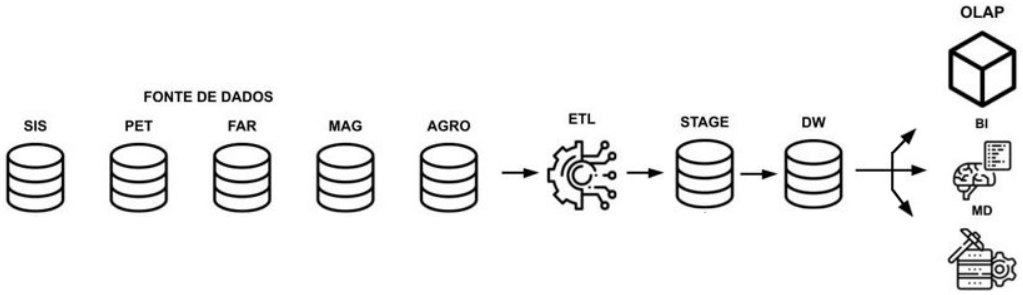
RESUMO: Este trabalho foi desenvolvido como parte de um projeto de inovação em parceria com uma empresa especializada em software para o varejo e se encontra na Fase II (segundo ano de execução). Na Fase I, destinada à construção de um *Data Warehouse* (DW), foram realizadas atividades de integração, análise, normalização e mineração de dados de empresas presentes no sistema da empresa. Nesta Fase II, buscamos aplicar as metodologias a dados de empresas de todos os segmentos, não só de Pet Shop como ocorreu na primeira fase, além de dar ênfase aos dados de compras das empresas. Este trabalho tem como objetivo classificar empresas por segmento, como por exemplo Pet Shop, Farmácia, entre outros. Para a realização dos objetivos, serão realizados diferentes experimentos com várias técnicas de mineração de dados, como o algoritmo de classificação *Naive Bayes*, que é baseado em treinamento e cálculo de probabilidade de acordo com esse treinamento. Nossos experimentos indicam que a aplicação de técnicas de comparação com o CNAE da empresa apresentaram 76% de eficiência na classificação, e concluiu com a aplicação do algoritmo *Naive Bayes*.

PALAVRAS-CHAVE: *Data Warehouse*; Mineração de Dados; Inteligência Artificial; Classificação.

INTRODUÇÃO

Em parceria com uma empresa especializada em software para o varejo, na primeira parte desse projeto foi feita a construção de um sistema envolvendo *Data Warehouse* (DW), *Business Intelligence* (BI) e Mineração de Dados (MD) voltado para vendas, e na segunda parte desse projeto fez essas etapas em relação as compras da empresa. Este trabalho tem foco na parte de MD, dentro da proposta de um projeto maior, composto por várias partes; cada parte está descrita em outros projetos, pelos respectivos discentes. A Figura 1 ilustra o projeto maior, com suas respectivas etapas.

FIGURA 1. Etapas presentes no projeto.



Fonte: Elaborada pelos autores.

A primeira etapa faz referência às diferentes fontes de dados. A empresa fornece sistemas ERP para diferentes áreas, como a de Comércio (SIS), Pet Shop (PET), entre outros. Cada cliente tem o seu próprio banco de dados, e o sistema de cada área pode usar tabelas específicas, mas todos os bancos compartilham de uma mesma estrutura em comum. A segunda etapa faz referência a carga desses dados na base operacional da empresa. Para fazer o processamento desses dados é necessário carregá-los em um mesmo banco de dados, o DW, mas para facilitar a normalização e carregamento eles são direcionados a *Stage*, e posteriormente carregados no DW. A terceira etapa faz referência a geração de relatórios. Com os dados normalizados e em um mesmo banco de dados, podemos fazer a geração dos relatórios por meio do BI, MD e OLAP.

Dentro da área de MD, está separada em duas áreas: a parte de classificação e de agrupamento, ilustrada na figura 2. A segunda etapa é referente à classificação da empresa presente na base de dados por segmento. Já a terceira etapa refere-se ao agrupamento de dados, buscando definir grupos de clientes e encontrar padrões e tendências para uma melhor tomada de decisão.

FIGURA 2. Área de Mineração de Dados.

Fonte: Elaborada pelos autores.

Com o intuito de identificar o segmento de cada empresa de forma automatizada, tendo em vista que os processos anteriores não realizam tal separação, faz-se necessária a utilização de Técnicas de Classificação, de maneira a impedir a comparação de dados de empresas de diferentes segmentos, como, por exemplo, comparar dados de um Pet Shop com dados de uma Farmácia. Neste trabalho, nos dedicamos ao problema de desenvolver um classificador automático capaz de identificar o ramo de atuação (segmento) de uma empresa.

FUNDAMENTAÇÃO TEÓRICA

Segundo Ponniah (2001), uma organização gera por semana o tanto de informação que uma pessoa pode ler em uma vida toda, tornando humanamente impossível o estudo e interpretação de todos os dados para encontrar as informações úteis. Portanto, para realização das transformações de dados das empresas, faz-se necessária a utilização de mineração de dados.

A mineração de dados é o processo de descoberta de conhecimentos que possibilitam compreender a substância dos dados de uma forma especial, desvendando padrões e tendências que seriam impossíveis apenas com os dados brutos.
Para a realização deste processo de descoberta de padrões e tendências, a mineração de dados segue uma sequência de passos, ilustrados na figura 3, que consiste em:

1. Determinar os objetivos do negócio**;**
2. Preparo dos dados**;**
3. Realizar a mineração de dados**;**
4. Avaliação, apresentação e incorporação de descobertas**.**

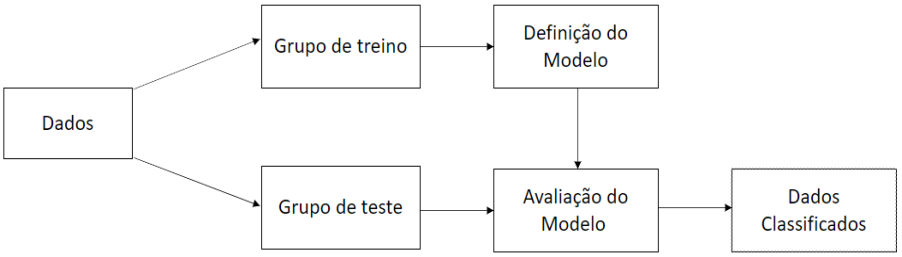
FIGURA 3. Passos para descoberta de padrões.

Fonte: Adaptado de Ponniah (2001).

Para a realização de tal classificação, faz-se necessária a utilização de algum algoritmo. Neste caso, o algoritmo escolhido foi o *Naive Bayes*, que segundo Garg (2013), é um algoritmo que assume cada parâmetro presente na classificação como independente, como exemplo a fruta "maçã", que tem diversas características que serão utilizadas como parâmetros, como sua cor, se tem semente, entre outros.

O algoritmo *Naive Bayes*, como ilustrado na figura 4, realiza a separação de dados que servirão como base para a classificação de novos dados, ou seja, um conjunto de dados pré-classificados para o grupo de teste. Este grupo de treino, que é o dos dados pré-classificados, é utilizado pelo algoritmo para definição de um modelo que posteriormente será aplicado no grupo de teste para definir o dado de acordo com as probabilidades geradas com a avaliação do modelo em conjunto com os dados de teste.

FIGURA 4. Algoritmo *Naive Bayes*.



Fonte: Adaptado de Navlani (2018).

METODOLOGIA

No desenvolvimento deste trabalho foi realizada uma série de experimentos, em busca de um método de classificação de empresas que apresentasse o melhor desempenho. Para a execução desses experimentos, a empresa parceira disponibilizou uma base de dados antigos, possibilitando a aplicação e desenvolvimento sem afetar diretamente a base de produção. Dentro desta base temos um total de 52 empresas, e dentre elas foram selecionadas 6 do segmento Pet Shop e 9 de outros segmentos para a realização dos primeiros experimentos.

Portanto, para começarem os testes, foram realizadas buscas dos dados que melhor definem o segmento da empresa, definindo então que os melhores a serem trabalhados são:

- Nome e Nome Fantasia da empresa;
- CNAE (Classificação Nacional de Atividades Econômicas) e CNAEs secundários, código que determina quais as atividades realizadas pela empresa, como exemplo o CNAE 9609-2/0, identificando a atividade "Higiene e embelezamento de animais domésticos".

Com os dados definidos, foi necessário uma busca por alguns deles, pois nem todos estavam presentes na base de testes, então foi utilizada a API Minha Receita[1] para obtenção dos dados que faltavam.

Após a busca por esses dados, foram identificadas as técnicas de classificação mais adequadas, dentro da linguagem de programação *Python*. Para isso então, as duas metodologias utilizadas foram:

- Classificação baseada em regras: análise comparativa entre termos pré-definidos;
- Aprendizado Supervisionado: *Naive Bayes*;

Com todos os dados, técnicas e algoritmos definidos, os testes começaram a ser executados. A primeira série de testes, que busca definir se a empresa é do segmento Pet Shop ou de outro segmento baseando-se no nome e nome fantasia da empresa, foi composta por 5 partes:

1. Análise comparativa entre os termos DOG e PET com o nome da empresa;
2. Análise comparativa entre os termos DOG e PET com o nome fantasia da empresa;
3. Análise comparativa entre vários termos referentes a Pet Shop com o nome fantasia da empresa;
4. Análise comparativa entre vários termos referentes a Pet Shop com o nome da empresa;
5. Aplicação da melhor análise comparativa em todos os dados da base de teste;

A segunda série de testes, agora buscando definir a empresa baseando-se nos CNAEs que compunham a empresa, combinando com a aplicação do *Naive Bayes*, foi realizada uma análise comparativa entre CNAEs das empresas com CNAEs definidos pelo IBGE de segmento Pet Shop, aplicando os resultados no algoritmo *Naive Bayes*.

RESULTADOS E DISCUSSÃO

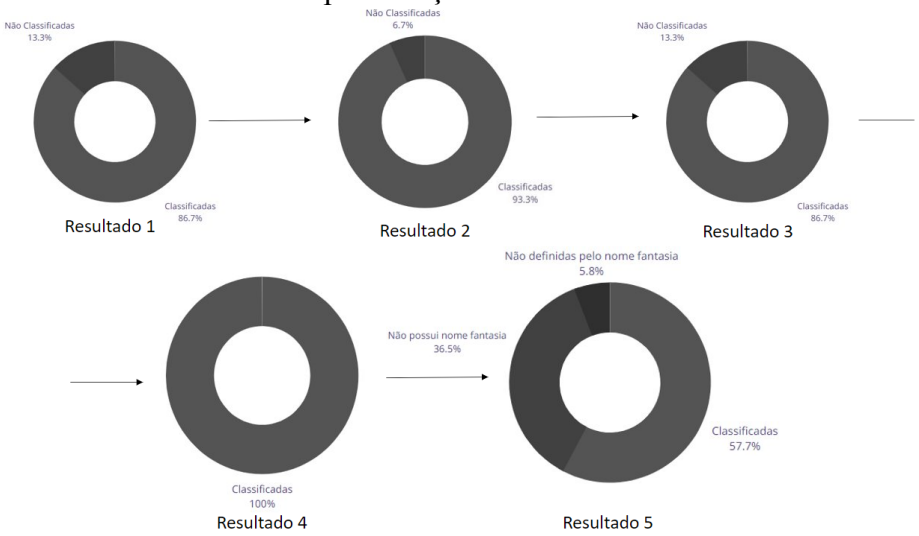
Dentro da primeira série de experimentos, assim como ilustrado na figura 5, tivemos os seguintes resultados:

1. Tivemos um total de 13 empresas classificadas corretamente pelos termos pré-definidos, e 2 não foram definidas, sendo ambas de segmentação Pet Shop;

[1] https://github.com/cuducos/minha-receita.

2. Tivemos um total de 14 empresas classificadas corretamente pelos termos pré-definidos, e 1 não foi definida, sendo ela de segmentação Pet Shop;
3. Tivemos um total de 13 empresas classificadas corretamente pelos termos pré-definidos, e 2 não foram definidas, não sendo nenhuma das duas de segmentação Pet Shop;
4. Tivemos um total de 15 empresas classificadas corretamente, 100% de precisão com os dados das empresas separadas para testes.
5. Tivemos um total de 30 empresas classificadas corretamente das 52 presentes da base, aproximadamente 57,7% de precisão.

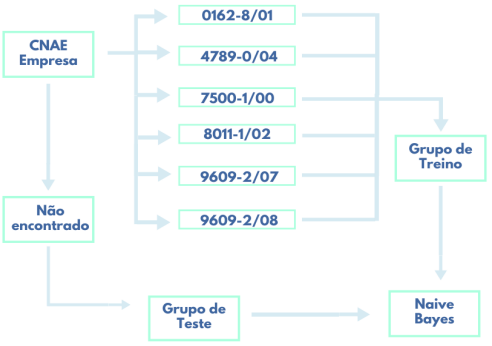
FIGURA 5. Apresentação dos resultados.



Fonte: Elaborada pelos autores.

Portanto, conclui-se que, apesar de 100% das empresas selecionadas serem classificadas, ao aplicar a mesma técnica em todas as empresas presentes na base de teste, o resultado não foi satisfatório, com isso, foi necessário a busca por novos métodos, pois o objetivo era uma precisão inicial superior a 80%. Com isto, utilizando como base os dados definidos anteriormente como melhores para classificação, foi efetuada uma nova série de experimentos, agora utilizado os CNAEs de cada empresa como elemento comparativo.

FIGURA 6. Classificação do CNAE.



Fonte: Elaborada pelos autores.

Após esta comparação inicial entre CNAEs da empresa e definidos como Pet Shop, tivemos um total de 39 classificadas corretamente de um total de 51 empresas, indicando uma porcentagem muito próxima de 80% (aproximadamente 76,47%), então, essas empresas classificadas serviram como grupo de treino para o algoritmo *Naive Bayes* para a definição do modelo, como ilustrado na figura 4, e o restante das empresas que não foram classificadas serão identificadas como grupo de teste ao algoritmo.

Após a identificação dos grupos de treino e teste, será aplicado o algoritmo *Naive Bayes* com a ideia apresentada anteriormente referente a comparação do nome da empresa, porém agora os termos pré-definidos serão buscados de forma automatizada, com o algoritmo definindo os termos que mais aparecem dentre os nomes das empresas. Após essa definição dos nomes que mais aparecem, cada termo será considerado um termo independente dentro do algoritmo, é comparado com o nome da empresa, caso encontrado o termo no nome, é identificado dentro de uma lista com o valor 1 na determinada posição do termo, e 0 para casos que não foram encontrados. Ao final é aplicado e comparado com o modelo gerado pelo *Naive Bayes* e determinado se a empresa é do segmento Pet Shop ou não.

CONCLUSÕES

Os processos de mineração de dados são necessários para uma definição mais eficiente do segmento da empresa, tendo em vista que torna os processos mais rápidos e eficientes se comparados à definição manual da empresa. Além disso, para que a empresa possa ter uma análise mais efetiva do mercado, e não se comparar com empresas de outros segmentos ao engano do sistema, as técnicas de classificação são fundamentais.

Após as discussões referentes à qual a metodologia seria mais efetiva para a classificação de uma série de empresas, e efetuando experimentos referentes a comparação entre o nome e nome fantasia da empresa com termos pré-definidos, e comparação entre CNAEs da empresa e pré-definidos de cada segmento, em conjunto com a aplicação do *Naive Bayes*, pôde-se concluir que os resultados para a tarefa de classificação foram satisfatórios, com cerca de 76% das empresas disponíveis na base de teste classificadas. No entanto, observamos que ainda há oportunidade para obtenção de melhorias. Nos nossos próximos passos, pretendemos aprofundar os experimentos com o algoritmo *Naive Bayes*, incluindo novas características (*features*) no modelo. Esperamos, assim, atingir níveis de precisão superiores a 90%.

REFERÊNCIAS

Garg, Bandana. "Design and development of naïve bayes classifier." (2013).

Navlani, Avinash. Naive Bayes Classification using Scikit-learn. Datacamp, 04 de dez. de 2018. Disponível em <https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>. Acesso em: 02 de set. de 2021.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 12, p. 2825-2830, 2011.

PONNIAH, P. Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. John Wiley and Sons, Inc, 2001.

Ribeiro, João Luiz Oliveira. "Uso de técnicas de mineração de dados em Python para classificação de pássaros baseado nas medidas dos ossos." (2017).

Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes a good classifier for document classification." International Journal of Software Engineering and Its Applications 5.3 (2011): 37-46.