Evento: XI Seminário de Inovação e Tecnologia

Adaptação de Algoritmos de Clusterização baseados em Machine Learning para o contexto da Pandemia de Covid-19

ADAPTATION OF MACHINE LEARNING-BASED CLUSTERING ALGORITHMS IN THE CONTEXT OF THE COVID-19 PANDEMIC

William Micael Hertz[1], Félix Hofmann Sebastiany[2], Sandro Sawicki[3]

RESUMO

O uso de Machine Learning e inteligência artificial vem crescendo em diversos segmentos juntamente com a evolução da arquitetura de computadores. Na área da saúde, por exemplo, é possível predizer tendências para determinadas doenças, ou ainda, auxiliar no tratamento e criação de novos medicamentos. No ano de 2020, com a pandemia mundial de Covid-19, o uso dessas tecnologias foi indispensável para prever e criar modelos de contaminação do vírus. Com base nessas informações, este projeto visa estudar algoritmos baseados em inteligência artificial, Birch e K-Means, para realizar futuras inferências entre base de dados relativa ao isolamento social e outros indicadores sociais como o CNAE (Classificação Nacional de Atividades Econômicas), com vistas a gerar informações sobre a evolução do Covid-19 no Estado do Rio Grande do Sul.

Palavras-chave: Covid-19, Inteligência Artificial, *Machine Learning*

INTRODUÇÃO

Em pleno século XXI é difícil imaginar a nossa vida sem a internet, pois dependemos dela para praticamente todas as tarefas, seja para trabalhar, estudar, fazer compras, entre muitas outras opções. O desenvolvimento de novas aplicações e a migração de outras já existentes para um contexto de computação em nuvem, tem se mostrado uma tendência crescente e irreversível ao longo dos últimos anos. Os avanços na tecnologia de computação em nuvem incluem um ecossistema de software que inclui aplicativos integrados, aplicativos implementados em máquinas virtuais em nuvem, aplicativos móveis, aplicativos

[1] Bolsista de Iniciação Científica, CNPq, aluno do curso de Ciência da Computação, Unijuí
[2] Aluno de Mestrado do Programa de Modelagem Matemática e Computacional da Unijuí
[3] Professor do Programa de Pós-Graduação em Modelagem Matemática e Computacional da Unijuí

de mídia social e muitos outros aplicativos e serviços na nuvem, mais conhecidos como *SaaS* (*Software as a service*).

Como resultado deste ecossistema heterogêneo com aplicações estruturadas e desenvolvidas com as mais diversas tecnologias, têm-se gerado um enorme volume de dados, que precisam ser gerenciados, processados e analisados, com o intuito de obter *insights* e métricas relevantes para a empresa. Uma destas ferramentas e técnicas que auxiliam é o "*big data*", refere-se a dados que são tão grandes, rápidos e complexos que são difíceis ou quase impossíveis de processar usando métodos tradicionais. O ato de acessar e armazenar grandes quantidades de informações para análise existe há muito tempo. Mas o conceito de big data ganhou impulso no início de 2000, quando o analista da indústria *Doug Laney* articulou a definição agora dominante de big data como os três V's: *Volume, Velocidade e Variedade* (SAS, 2021).

Com todos estes dados à disposição, é necessário ferramentas capazes de tomar decisões e resolver problemas, aí surge a inteligência artificial. A IA (Inteligência artificial) é onipresente hoje, usada para recomendar produtos em lojas, assistentes virtuais, reconhecimento facial, análise e crescimento de processos de fabricação entre muitos outros.

No ano de 2020, com a pandemia mundial de Covid-19 (SARS-CoV-2), o uso de *Machine Learning* e *Deep Learning* auxiliaram diversas pesquisas na criação de modelos de contaminação do vírus (REHBEIN, 2020). Neste sentido, este resumo expandido mostra como está sendo elaborada a infraestrutura computacional para correlacionar variáveis e bases de dados distintas com vistas a inferir conhecimento usando algoritmos de clusterização baseados em *Machine Learning, Birch* e *K-Means* no contexto da Pandemia de Covid-19.

METODOLOGIA

A pesquisa inicialmente foi exploratória com o objetivo de conhecer as ferramentas, técnicas, bibliotecas e tecnologias, após, foram realizados testes práticos por meio de códigos fonte *Python*.

Foi estudada e utilizada a base de dados do CNAE, que é a Classificação Nacional de Atividades Econômicas usada para identificar as funções e obrigações das empresas em todos os municípios brasileiros, com o objetivo de isolar a variável "vínculo empregatício". Além disso, foram interpretados os dados de casos diários de Covid-19 por municípios do RS

(Rio Grande do Sul) disponibilizados pela Secretaria da Saúde do Estado do Rio Grande do Sul. Na sequência foram utilizadas bibliotecas para a aplicação de *Machine Learning*, como a *pandas*, *sklearn* e a *matplotlib*. Com isso, foram estudados e adaptados os algoritmos de clusterização *Birch* e *K-Means* visando a criação de classes juntamente com a correlação com a variável "vínculo empregatício".

RESULTADOS E DISCUSSÃO

Obrigatória a todas as pessoas jurídicas, inclusive autônomos e organizações sem fins lucrativos, a CNAE (Classificação Nacional de Atividades Econômicas) é essencial para obtenção do CNPJ (Cadastro Nacional da Pessoa Jurídica). Além de contribuir para melhorar a gestão tributária do país, também é possível fazer um levantamento de qual a atividade econômica mais exercida em um determinado município, assim como também é possível saber os vínculos empregatícios existentes.

Utilizando a linguagem *Python* e algumas bibliotecas utilizadas no desenvolvimento de machine learning, como a *pandas, sklearn* e a *matplotlib*, se tem um código para analisar os casos de coronavírus a cada 100 mil habitantes. Com a biblioteca *Pandas* é feito um "*for*" para ler 21 arquivos com datasets do CNAE (Classificação Nacional de Atividades Econômicas), onde contém as informações dos municípios. Com a criação de clusters é possível realizar o tratamento dos dados que existem no CNAE (Classificação Nacional de Atividades Econômicas).

Posteriormente modificado para alterar os tipos de métodos de agrupamento utilizados e também os grupos obtidos anteriormente, que são 21, utilizando os métodos de *Birch* e *KMeans*.

Utilizando o método *Birch*, o grupo econômico que teve os maiores números de infectados por 100 mil habitantes foram os que têm predominância em administração pública, sendo da saúde e do comércio e atividades essenciais, e a menor taxa foi no grupo com predominância de atividades financeiras e serviços profissionais e técnicos.

Já com o método do *K-Means*, os maiores índices foram nos grupos de administração pública, e a classe industrial apresentou um crescimento semelhante ao comércio e atividade essenciais. E com os menores índices de crescimento apresenta predominância Transformação/Industrial e, Comércio e atividades essenciais, enquanto que a classe 6 possui

sua predominância no grupo econômico de Comércio e atividades essenciais, e Administração pública. Ou seja, ambas as classes apresentam altos níveis de Comércio e atividades essenciais.

CONSIDERAÇÕES FINAIS

Dos 21 grupos originais, foram criados 7 grupos, os quais foram aplicados os algoritmos de clusterização *Birch* e *K-Means*. Ambas as técnicas apresentaram clusters similares em seus resultados. Com base em modelos estatísticos e matemáticos foi possível encontrar alguns dados preliminares. Grupos das atividades econômicas relacionadas ao setor agropecuário, ao comércio e atividades essenciais representam uma maior quantidade de cidades. As classes que tiveram a maior quantidade de casos por cem mil habitantes foram as classes de Comércio e Atividades Essenciais e, também, a classe da Saúde.

Além disso, o uso de ferramentas baseadas em *Big Data* e Machine Learning vem crescendo, mostrando que o mercado de trabalho está mudando, alterando a forma de como as pessoas se comportam em relação às máquinas, como as automações de processos e até mesmo por algumas vezes, a substituição da mão de obra humana por máquinas.

AGRADECIMENTOS

Agradeço ao professor orientador Prof. Dr. Sandro Sawicki por sempre ter me ajudado e pela oportunidade de aprofundar os meus conhecimentos nestes assuntos. Ao Félix, novo amigo que encontrei ao decorrer dos estudos. SARS-CoV-2. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, através do Programa Institucional de Bolsas de Iniciação Científica da Unijuí que foi de enorme importância para aquisição de conhecimento e crescimento intelectual.

REFERÊNCIAS BIBLIOGRÁFICAS

REHBEIN, Matheus H.. **Comparação de Métodos Não Supervisionados: Um Caso Baseado no CNAE 2.0 e na Covid-19**. 2020. 13 f. Relatório Técnico. Programa de Pós

Graduação em Modelagem Matemática e Computacional, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2020.

SAS (ed.). **Big Data: what it is and why it matters. What it is and why it matters**. Disponível em: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html. Acesso em: 02 jul. 2021.

VERSIANI, Rafael. **CNAE: o que significa e qual a importância para o seu negócio**. Disponível em: https://enotas.com.br/blog/cnae-o-que-e/. Acesso em: 01 jul. 2021.