Plan Overview

A Data Management Plan created using DMPTool

Title: Ensembles de Algoritmos de Classificação - Árvores de Decisão, Naïve Bayes e k-NN - em Mineração de Dados de Exames de Pacientes no Diagnóstico do Covid-19

Creator: Samuel oliveira Da silva

Affiliation: State University of Campinas (unicamp.br)

Principal Investigator: Samuel Oliveira da Silva

Project Administrator: Samuel Oliveira da Silva, Alan Gonçalves

Template: UNICAMP-GENERICO: Aplicável a todas as áreas

Project abstract:

Estes dados foram coletados para um estudo sobre o uso de Ensembles de Algorítmos de classificação em mineração de dados de Exames de Pacientes no Diagnóstico do Covid-19. Abordamos assuntos relacionados à Técnicas de Seleção de Atributos, Identificação e Tratamento de Outliers, Balanceamento de Classes, Correlação de Atributos, Limpeza e Transformação de dados, Tratamento de Dados Faltantes, além de mencionar os algorítmos de Classificação Decision Tree, Naïve Bayes e k-NN. Finalmente mostramos a aplicação de uma Matriz de Confusão e resultado da acurácia após treinar os modelos.

Start date: 01-01-2021

End date: 06-30-2021

Last modified: 06-21-2022

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

---

Ensembles de Algoritmos de Classificação - Árvores de Decisão, Naïve Bayes e k-NN - em Mineração de Dados de Exames de Pacientes no Diagnóstico do Covid-19

Descrição dos Dados e Metadados

Quais serão os dados coletados?

O dados coletados são relativos a pacientes que foram atendidos no Hospital Adventista de Manaus, e realizaram exames no laboratório do hospital, no periodo de 01/01/2021 a 31/05/2021.

Contém os resultados de exames laboratoriais rotineiros realizados em pacientes clínicos e cirurgicos, todos os pacientes realizaram o exame especifico para COVID-19(**rt_pcr**), e outros exames relacionados ao acompanhamento do paciente.

Sãos os exames:

- rt_pcr
- leucocitos
- basóficos
- proteina_c
- hemoglobina

São dados:

- Numericos
- Categóricos

Representam os resultados dos exames realizados. Não havendo relação entre a quantidade de registros contendo os resultados dos exames e a quantidade de pacientes atendidos, pois um paciente pode ter realizados os mesmos exames mais de uma vez ao longo do período do levantamento. São exames para acompamento da evolução clínica ou do quadro pós-cirúrgico do paciente.

Que metadados serão anotados e qual padrão será seguido?

As amostras são de pacientes que foram admitidos para internação com suspeita de Convid-19 e foram submetidos a exames que seguem o protocolo estabelecido de diagnóstico adotado naquela instituição, além do exame **rt_pcr** ***(covid-19)***, também sistematicamente eram realizados outros exames que dão suporte ao diagnóstico do Covid-19.

Ao padrão de armazenamento dos dados é o **C**omma-**S**eparated-**V**alues (CSV) que é um arquivo texto **separados ou delimitados por uma vírgula.** A primeira primeira linha corresponde ao nome dos atributos, as linhas seguintes são os registros ou as linhas com os valores dos atributos.

Atributo	Descrição	Natureza
idade	Idade do Paciente	Numérico
rt_pcr	Exame Covid-19	Categórico
leucócitos	Exame Laboratoriais	Numérico
basofilos	Exame Laboratoriais	Numérico
creatinina	Exame Laboratoriais	Numérico
proteina_c	Exame Laboratoriais	Numérico
hemoglobina	Exame Laboratoriais	Numérico

Aspectos Legais e Facilidade de Acesso aos Dados

Quais são as questões legais e éticas associadas aos dados e relevantes a este projeto?

Todos os dados disponibilizados por este estudo deverão ser adequadamente referenciados.

Os dados poderão ser utilizados, como forma de otimização do uso de recursos.

Os dados não contam com restrições quanto ao uso de seres humanos ou animais, propriedade intelectual, nem são dados reaproveitados de outros estudos.

Quais são as políticas a serem utilizadas para o compartilhamento de dados?

Não existem questões éticas ou jurídicas que requeiram prévia atenção do compartilhamento dos dados obtidos.

Os dados foram gerados e cedidos pelo Hospital Adventista de Manaus, omitindo os nomes ou qualquer forma de identificar os pacientes que realizaram os exames.

São dados crus e poderão ser compartilhados publicamente, sem a necessidade de solicitações. Em contrapartida, ao utilizar os dados disponibilizados, a parte interessada deverá referenciar o local onde os dados estarão armazenados, bem como os pesquisadores responsáveis.

Os dados serão armazenados no Repositório de Dados de Pesquisa ZENODO, que disponibiliza a maneira requerida para citação dos dados.

Gestão de Dados e Armazenamento

Em que formatos serão armazenados os arquivos resultantes da pesquisa em questão? Que software poderá ser utilizado para a manipulação de cada um dos formatos listados?

Os dados numérios e categoricos serão disponibilizados em arquivos no formato CSV e poderão ser visualizados utilizando planilhas eletrônicas como Microsoft Excel, Libre Office e um grande número de ferramentas, incluindo Notepad, Vi, Emacs e TextEdit entre outros.

Como e onde estes arquivos serão mantidos? Por quanto tempo ocorrerá esta preservação? Como será realizado o backup destes dados?

Os dados serão armazenados no Repositório de Dados de Pesquisa ZENODO. Onde será mantido por tempo não definido (limitado às condições do repositório) e versionado conforme os recursos disponíveis pela ferramenta.

O versionamento aplicado ao dataset é uma adaptação do Versionamento Semântico (SemVer), as informações estão diponíveis no site https://semver.org/lang/pt-BR/.

O acesso aos dados será realizado por meio do DOI https://doi.org/10.5281/zenodo.6675015.