DEFESA DE TEMA

Clustering Machine Learning aplicado a Pacientes Pertencentes ao Espectro Autista

Israel José Monteiro Carvalho
israeljmcarvalho@gmail.com

Católica de Santa Catarina
Graduação em ENGENHARIA DE SOFTWARE
Disciplina: T2ESOFT08N | PORTIFÓLIO DE PROJETO

Católica de Santa Catarina
Centro Universitário

O projeto

- **Quem é o cliente?**

Como potenciais clientes deste projeto destacam-se profissionais da área médica, psicologia, Terapia Ocupacional, Fonoaudiologia, dentre outras áreas onde são realizadas estudos relacionados a pacientes pertencentes ao espectro autista

O projeto

- **Quais os problemas ou oportunidades temos para resolver?**

Devido a falta de informações sistematizadas, os profissionais de área da saúde que cuidam de pacientes do TEA tem dificuldades de classificar seus pacientes incorrendo geralmente em ponderações arbitrárias para tal clusterização, o que é sabido ser algo plenamente impreciso.

Pleiteia-se criar um modelo de Machine Learning onde será utilizado o algoritmo não supervisionados chamado Custering. É fruto ainda deste trabalho criar, parametrizar, e executar este algoritmo na linguagem R.

O projeto

- **Qual o benefício claro que o cliente pode ter?**

É muito comum analisarmos observações dentro de um contexto (ou grupo) afim de identificarmos padrões de comportamentos dos registros observados. Na área médica, não é diferente. Hoje, médicos e demais profissionais da área da saúde agrupam pacientem com base em seu conhecimento e observações subjacentes, mas diante de dezenas (ou até centenas) de variáveis de uma paciente, é inevitável que ocorre ponderação arbitrária atribuindo valores de forma empírica (consequentemente não científica) na separação de pacientes em grupos afim de poder estudar estes pacientes sob a ótica do grupo onde está inserido.

Espera-se que com o resultado apresentado por este algoritmo, os profissionais da saúde tenham mais precisão ao analisar possíveis grupos de indivíduos visto que estes grupos (clusters) emergirão fruto deste poderoso algorítimo de Machine Learning, o Clustering.

O projeto

- **Como será a experiência do cliente nesse novo serviço?**

Em um primeiro momento, como fruto deste trabalho, serão ofertados aos clientes da área de saúde uma relação de grupos com seus respectivos pacientes oriundos da clusterização do algoritmo de *Clustering* sem que haja quaisquer ponderações arbitrárias. O número de grupos poderá ser definido pelo cliente / restrição do problema de negócio (*k-means method*), ou ser "sugerido" pelo método *Hierarchical Cluster Analysis*. O cliente participará desta decisão pois a escolha do método influenciará diretamente o resultado.

Posteriormente, pode-se avaliar a criação de uma interface a nível de usuário para coleta das informações de cada paciente, mas como fruto deste projeto sugiro foco na parte de DataScience (Machine Learning) sendo disponibilizado uma tela desenvolvida em Python onde o cliente escolherá o dataset bem como dará start no modelo.

O projeto

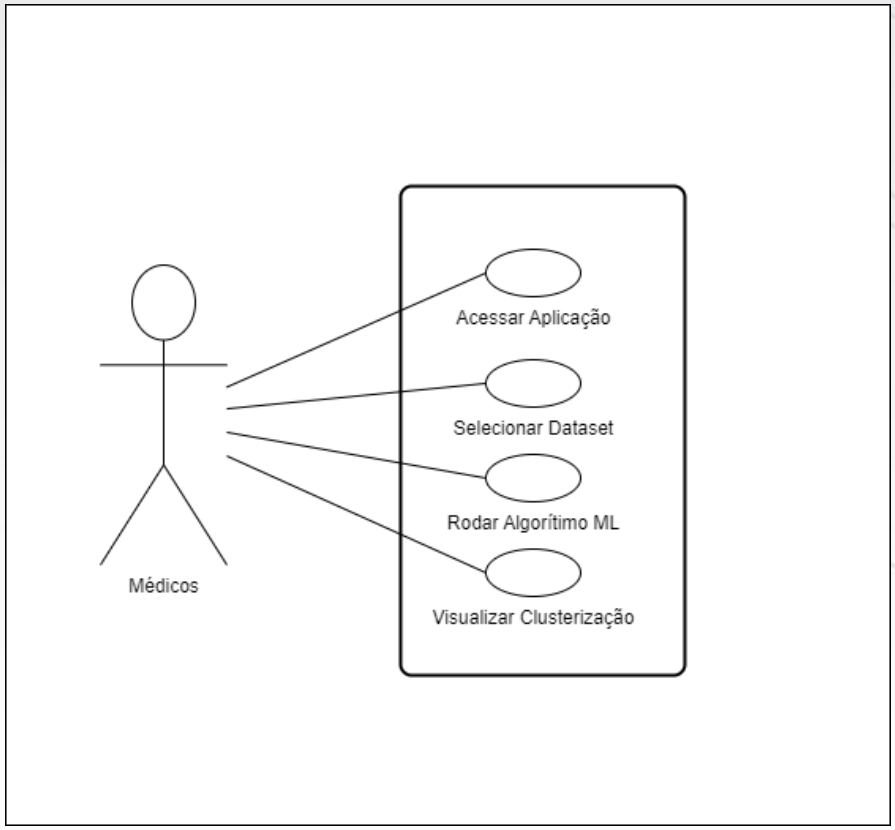
- **Qual o benefício claro que o cliente pode ter?**

Os clientes terão acesso a classificação de seus pacientes com base em critérios científicos (*euclidean distance*) e não em ponderação arbitrária.

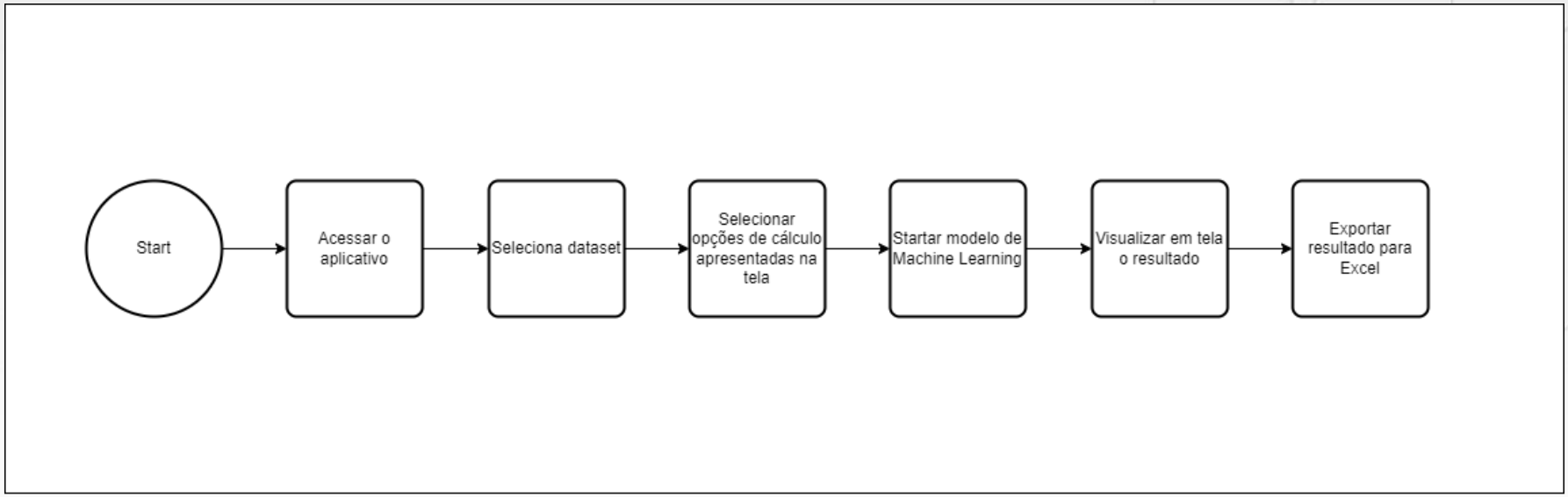
Cita-se ainda que o algorítimo de clustering é extremamente performático (em datasets com tamanho apropriado para a área da saúde) bem como altamente escalável e replicável.

Casos de Uso

Diagrama de casos de uso



Flowchart


Start
Acessar o aplicativo
Seleciona dataset
Selecionar opções de cálculo apresentadas na tela
Startar modelo de Machine Learning
Visualizar em tela o resultado
Exportar resultado para Excel

Requisitos

Requisitos Funcionais

- Upload de Dataset para realização da clusterização
- Selecionar o método de calcular para as distâncias das observações (Euclidean Distance, Minkowski distance, Chebychev distance, Canberra distance ou Manhattan distance)
- Selecionar o método de encadeamento do algorítimo (Complete Linkage, Single Linkage, Centroid Method)
- Executar algorítimo de Clustering
- Apresentar para usuário resultado da clusterização

Requisitos

Requisitos Não Funcionais

- Tela de fácil entendimento
- Rápida parametrização
- Algorítimo deve ser rodado em R
- Algorítimo deve ser performático

Pacotes do Feature Driven Development (FDD)

Pacote 1 – Data 11/09/2023

- Estrutura central do projeto contendo definições de clientes, objetivo do trabalho, problemas ou oportunidades, requisitos, diagramas e metodologia a ser utilizada

Pacote 2 - Data 02/10/2023

- Algorítimo de Machine Learning Clustering já em funcionamento a partir de scripts em R

Pacote 3 - Data 23/10/2023

- Interface gráfica já se comunicando com o código em R

Pacote 4 - Data 30/10/2023

- Aplicação rodando no Heroku

Pacote 5 - Data 06/11/2023

- Documentação Técnica (Readme.md) bem como fundamentação teórica do trabalho

Tecnologias aplicadas

- Linguagem
  R / Python
- Banco de Dados
  Dataset Local
- Ferramentas
  Rstudio / Pycharm
- bibliotecas

  R

  library(tidyverse) / library(cluster) /library(dendextend) /library(factoextra) / library(fpc) / library(gridExtra) / library(readxl) / library (reshape) / library(dplyr)

  PYTHON

  Django / Pandas