Fatec São José dos Campos
Prof. Jessen Vidal

CENTRO PAULA SOUZA

GOVERNO DO ESTADO SÃO PAULO

Trabalho Prático - EpidemiWeb

Arthur Barbero;
Felippe Alves;
Gabriel Landim;
José Vinicius Santana;
Thyago Odorico.

Inteligência Artificial - Análise e Desenvolvimento de Sistemas

Prof. Me. José Walmir G. Duque. – 6º semestre – 2021

Sumário

1. Introdução

2019 foi um ano que irá ficar nos arautos da história, principalmente pela grande pandemia de uma nova doença com origem nos países asiáticos, uma evolução da SARS, denominada COVID-19.

A ideia de passar novamente por uma epidemia assusta qualquer pessoa nos dias atuais, e olhando as ações que foram tomadas, a única forma de minimizar os problemas e causas é a rápida identificação e comunicação, ou seja, informação.

Considerada como o novo petróleo, a informação é tida como o grande recurso dos dias atuais, e não são poucos os tipos de softwares que tentam acessar e coletar nossos dados dia após dia na esperança de vincular nossos dados ao consumo ou estilo de vida, para que as empresas possam oferecer seus serviços ou despontar à frente de seus concorrentes. Com essa visão, a informação também pode ser usada para o bem das sociedades, basta que consigamos reter informações dos usuários de forma inteligente, concisa e transparente.

Desta maneira, utilizaremos formas de capturar informações, retê-las e utilizá-las de forma a identificar novas doenças, padrões de relacionamento entre doenças-pacientes e doenças-sintomas, padrões de localidade e gerar insights de possíveis novas epidemias.

A Inteligência Artificial entra neste cenário auxiliando o projeto como um todo e se beneficiando de seus dados que serão coletados ao longo de seu uso visando a implementação de modelos treinados à uma aplicação WEB.

2. Business Understanding

2.1. Objetivo

É de conhecimento geral que o setor da saúde, na maioria dos países, é majoritariamente privado e de acesso a poucos. No Brasil e em outros poucos países como Reino Unido, Canadá, Dinamarca, Suécia, Espanha entre outros, o acesso a saúde, de forma unificada e universal, é um direito do cidadão. Restringindo o escopo para o território brasileiro, podemos identificar que o SUS (Sistema Único de Saúde) é o grande catalizador de informações relacionadas a ocorrência de doenças e mantêm histórico que pode ser acessado em qualquer Estado brasileiro. Segundo o IBGE em 2019, 71,5% da população brasileira dependem exclusivamente do SUS para diversos tratamentos.

Pensando no grande fluxo de dados públicos que poderíamos utilizar e na transparência de dados disponibilizados ao público, seria de fácil acesso para que pudéssemos traçar ligações entre doenças e sintomas aos locais de ocorrência, ou até com informações socioeconômicas, porém, infelizmente os dados que podemos acessar são escassos, e por mais que tenhamos acesso as estas informações do Portal da Saúde, só podemos visualizar dados de estatísticas vitais ou de mortalidade, algumas doenças epidemiológicas são citadas mas não existe a possibilidade de verificar novas doenças ou possíveis epidemias pela falta de padrão e informações mais qualificadas como localização, datas exatas, gênero, idade e etc ...

Com este objetivo, coletando as informações de forma padronizada e com máximo de aproveitamento, utilizaremos a Inteligência Artificial para criar modelos que realizam diagnósticos preliminares com base nos sintomas informados pelo usuário.

2.2. Funcionalidades e Requisitos do projeto

Conforme o proposto, os principais atores do projeto são:

Papel	Descrição	Nível de acesso
Agente da saúde	Maior responsável ao acesso e inserção de informações pertinentes aos cidadãos que são atendidos no estabelecimento da saúde como também pelas informações de incidência da doença e sintomas descritos pelos mesmos.	Tático
Cidadão	Usuário que deseja o acesso as informações coletadas e disponibilizadas com transparência, dentro das normas previstas em lei, podendo consultar os dados e fazer uso dos modelos treinados para identificar possíveis epidemias em sua região como a realização de diagnósticos preliminares de doenças a partir dos sintomas informados	Operacional

Dado os principais atores do projeto, para que seja possível a conclusão do projeto, o mesmo deverá atender os seguintes requisitos:

Ator	Tarefa	Regra de negócio
Agente da saúde	Cadastrar Doença	R1 – A doença deverá conter os campos “Nome” e “Data de criação”;
Agente da saúde	Cadastrar Sintoma	R1 – O sintoma deverá conter os campos “Nome”, “Descrição”, “Severidade” de 1 a 5 e “Data de criação”; R2 – Ao criar um sintoma, o mesmo deve ser relacionado ao menos a uma doença já cadastrada;
Agente da saúde	Atribuir Sintoma	R1 – A partir dos sintomas existentes, o agente poderá relacionar um sintoma a outra doença existente;
Agente da saúde	Cadastrar Incidência	R1 – No atendimento de um paciente, sua incidência deve ser criada para registro do acontecimento de uma nova doença; R2 – Para a criação da incidência, deverá conter relação com uma doença já existente e a um usuário já existente; R3 – Também serão necessários o preenchimento dos campos “Data da incidência” e “Data de criação”;

Agente da saúde	Cadastrar Usuário	R1 – Será necessário o preenchimento dos campos “Nome completo”, “E-mail”, “Senha”, “Endereço”, “Número do endereço”, “Bairro”, “Cidade”, “Estado”, “País”, “Data de criação”; R2 – Usuários com permissão de “Agente da saúde” poderão escolher também qual o perfil do usuário que está sendo criado.
Qualquer usuário	Diagnóstico Preliminar	R1 – Com a escolha dos sintomas já existentes e sua respectiva severidade, de 1 a 5, demonstrar a doença que preliminarmente se aproxima do caso apontado. R2 – Demonstrar também a porcentagem de proximidade de cada sintoma escolhido com a ocorrência da doença.

O projeto também deverá ser de fácil implementação e utilização, podendo ser acessado via WEB e também via integração nos padrões de uma aplicação REST.

2.3. Fases do projeto

- **Entendimento dos dados**
  Após o entendimento do negócio, necessitamos entender a disposição dos dados que iremos coletar e realizar as relações necessárias entre doença, incidências, sintomas e usuários, para que assim possamos construir um banco de dados na qual o projeto possa acessar para criar seus modelos e treina-los;

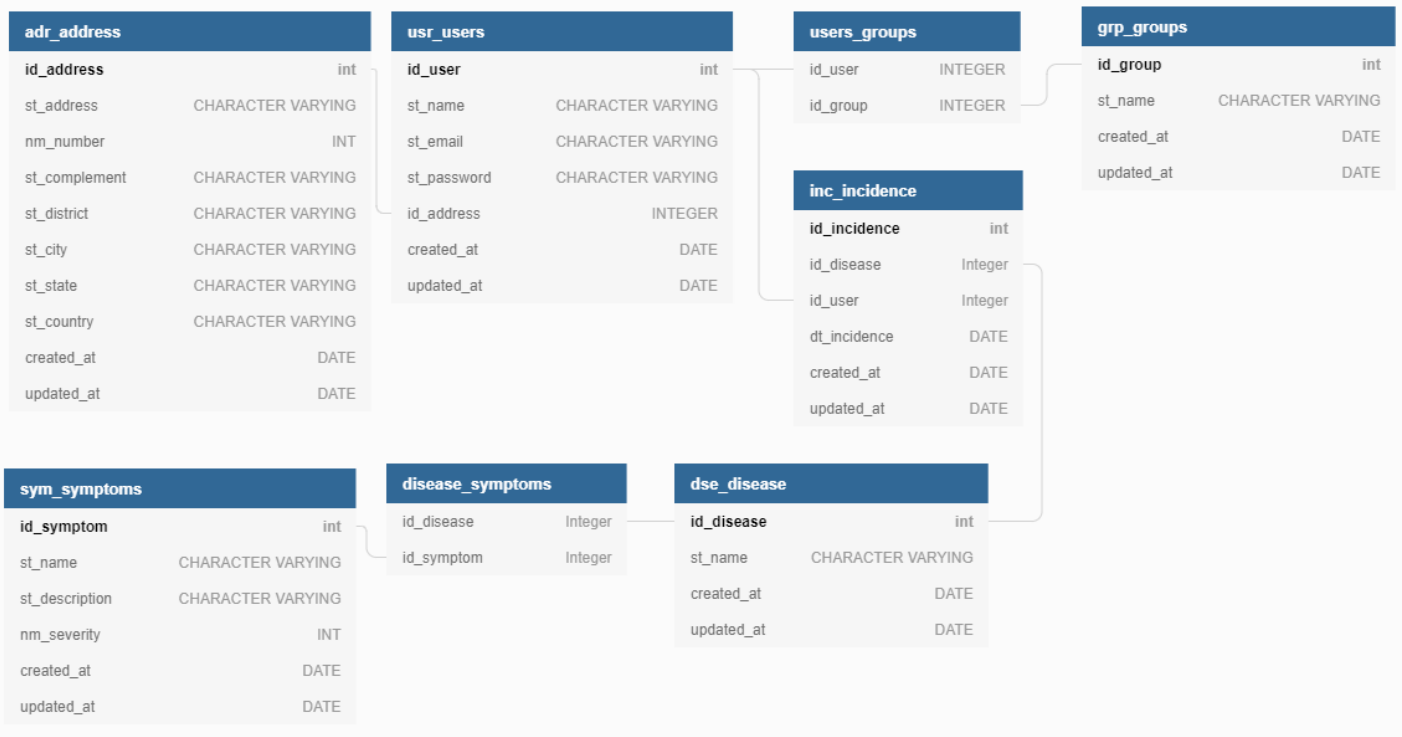
- **Confecção de Bancos de Dados**
  Com base nos requisitos e atores, criar as entidades e relacionamentos entre tabelas para que o projeto seja suportado por uma padronização e concentração de dados;

- **Criação dos modelos e rotas**
  Realizar a criação dos métodos que irão realizar o treinamento e iniciação de modelos como também as rotas que irão ser consumidas para a entrega dos resultados via API.

3. Data Understanding

Para atingirmos os objetivos do projeto, faz-se necessário um entendimento mais aprofundado dos dados, as disposições de seus relacionamentos e a identificação entre subgrupos, qualidade, relevância entre outros.

A criação de um banco de dados para o recebimento dos dados coletados do uso da ferramenta foi concebido no seguinte diagrama de entidade relacionamento:



Armazenamos os dados dos usuários entre 3 tabelas, sendo elas “usr_users” que possui a informação de nome, e-mail e senha, utilizados para entrar na plataforma, uma relação um para um com a tabela “adr_address” que armazena todas as informações de endereço do usuário e a tabela “grp_groups” que armazena as permissões do usuário.

Já os dados da regra de negócio são armazenadas entre as tabelas “dse_disease” que possuí o nome da doença, com relação muitos para muitos com a tabela “sym_symptoms”, que armazena os sintomas e sua severidade sobre a doença relacionada, também possuímos a tabela “inc_incidence” que armazena as incidências das doenças, relacionando uma doença à um usuário.

Com o entendimento das relações entre as informações, necessitamos agora entender como o projeto irá se utilizar destes para conseguir realizar os diagnósticos preliminares. Entre os dados possuímos as seguintes variáveis:

- Symptom (Sintoma):
  O sintoma é um atributo nominal, que representa um sintoma do paciente;

- Severity (Severidade):
  Atributo numérico que representa a severidade de um sintoma sobre uma doença em específico;

- Disease (Doença):
  De tipo nominal, o atributo da doença representa o diagnóstico na qual o paciente pode estar sendo acometido.

Tendo em vista que o especialista no domínio é a pessoa mais qualificada para realizar uma avaliação de peso entre o sintoma e a doença, seu input decorrente das ocorrências ao longo da vida do projeto irão alimentar o "dataset" com mais informações a cada incidência, conseguindo assim fazer predições preliminares das doenças a partir dos sintomas e pesos listados.

Exemplo da disposição dos dados:

Symptom_id	Disease_id	Severity	Symptom_desc	Disease_desc
1	163	2	Dor abdominal superior	Inflamação da colicistite da vesícula biliar
4	20	1	Abuso de álcool	Intoxicação por álcool etanol
5	732	1	Ansiedade (nervosismo)	Estresse
6	729	2	Dor ou aflição no braço	Tensão muscular puxando músculo

Utilizando uma abordagem de Machine Learning do tipo "Não-Supervisionado", teremos como variável de objetivo o atributo "Disease_desc" informando ao paciente o possível diagnóstico.

4. Data Preparation

Dado as informações sobre o “dataset”, necessitamos preparar os dados utilizando técnicas de “Data Mining” que mais se adequam ao problema principal do projeto, o diagnóstico preliminar de doenças dada seus sintomas.

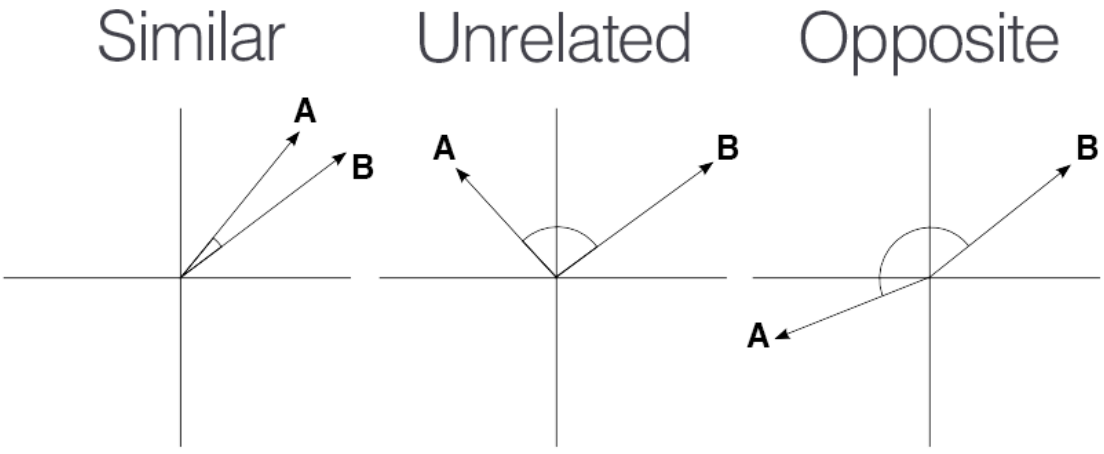
4.1. Tarefas de Machine Learning empregadas

As possíveis tarefas a serem utilizadas nesse processo são:

- **Associação**
  A Associação visa identificar as relações dentre os atributos dos dados e geralmente apresentam a forma “SE *atributo* X ENTÃO *atributo* Y”. é uma das tarefas mais conhecidas no mercado devido aos ótimos resultados no varejo, onde é sugerido itens compatíveis com uma lista de compras ou carrinho de compras que o usuário já esta inclinado a realizar a compra.

  Em nosso projeto utilizamos a associação em cima de uma regra estatística denominada **Okapi BM25 (Best Match 25)** similar a TFIDF (Term Frequency, Inverse Document Frequency), ela determina a frequência em que um atributo ou termo aparece dentro de um documento ou coleção de dados, representando esta frequência em uma escala de -1 até 1, onde -1 significa mais distante e 1 mais similar o atributo é dado sua coleção. Esta associação servirá como peso para nossa segunda tarefa de “Data Mining”.

- **Agrupamento (Clustering)**
  O agrupamento é uma tarefa de “Data Mining” que visa identificar e aproximar os registros similares entre si. Dentre as várias técnicas utilizadas para avaliar as similaridades entre os atributos, para este projeto utilizaremos a “Cosine Similarity” (Similaridade por cosseno), onde dentre um intervalo fechado [-1,1], é medida a similaridade entre dois vetores em um espaço vetorial.



Em decorrência do “dataset” utilizado, optamos por utilizar estas duas tarefas de “Data Mining” por se tratar de aprendizados não-supervisionados e também pela utilização dos pesos e da incidência dos sintomas e das doenças. Conforme mais incidências ocorrerem, a atribuição de pesos pelo algoritmo BM25 irá se atualizando e sua similaridade, atribuída pelo algoritmo de “Cosine Similarity”, identificará de forma mais assertiva quais doenças estão mais próximas aos sintomas informados.

4.2. Preparação dos dados

Realizamos as importações necessárias e a leitura dos dados removendo quaisquer valores nulos possíveis e forçamos o tipo das colunas de sintomas e doenças para “category” que converte tipos de número ou texto para itens únicos, salvando um pouco de memória dos processos.

```
import numpy as np
import pandas as pd
import matplotlib as plt
from scipy.sparse import coo_matrix
from scipy.sparse.linalg import svds
from sklearn.metrics.pairwise import cosine_similarity

data = pd.read_csv('archives/dataset.csv', ';')
all_symptoms = pd.read_csv('archives/all_symptoms.csv', ';')
all_diseases = pd.read_csv('archives/all_disease.csv', ';')

data.head()
```

	Symptom_id	Disease_id	Severity	Symptom_desc	Disease_desc
0	1	163	2	Upper abdominal pain	Cholecystitis inflammation of the gallbladder
1	1	164	2	Upper abdominal pain	Choledocholithiasis stone in bile duct
2	1	165	1	Upper abdominal pain	Cholelithiasis gallstones
3	1	187	2	Upper abdominal pain	Constipation
4	1	306	2	Upper abdominal pain	Gastric ulcer stomach ulcer

A partir destes dados removemos as colunas que não serão necessárias para os processos e criamos uma matriz esparsa em cima das informações de severidade, sintomas e doenças, a fim de economizar espaço eliminando as informações sobre cada coluna transformando o “dataframe” em uma lista de vetores contendo apenas seus valores normalizados entre cada item distinto.

```
data = pd.read_csv('archives/dataset.csv', ';', usecols=[0,1,2], names=['symptom', 'disease','sev'],
skiprows=1)

data = data.dropna(axis=0, how='any')  #drop nan

# map each disease and symptom to a unique numeric value
data['symptom'] = data['symptom'].astype("category")
data['disease'] = data['disease'].astype("category")

# create a sparse matrix of all the symptoms/severities
matrix = coo_matrix((data['sev'].astype(float),
                     (data['disease'].cat.codes.copy(),
                      data['symptom'].cat.codes.copy())))

sum_disease = data.groupby(['disease']).sev.sum()
symptoms = data['symptom'].cat.codes.copy()
```

Com os seguintes dados, estamos prontos para a utilização dos algoritmos de Associação, **BM25**, e de Agrupamento, **Cosine_similarity**, :

- "**data**":

	symptom	disease	sev
0	1	163	2
1	1	164	2
2	1	165	1
3	1	187	2
4	1	306	2

- "**matrix**":

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 1., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

5. Glossário

- Dataset – Coleção de dados;

- Machine Learning – Aprendizado de máquina;

- Data Mining – Mineração de dados;

- Dataframe – Conjunto de dados tabulados em duas ou mais dimensões.

6. Referências Bibliográficas

"Symptom Disease sorting" Dataset utilizado, Kaggle (https://www.kaggle.com/plarmuseau/sdsort) último acesso em 18/04/2021;

"Mineração de Dados Utilizando Aprendizado Não-Supervisionado: Um estudo de caso para bancos de dados da saúde" Campos Serra Domingues, Miriam Lúcia (https://www.lume.ufrgs.br/bitstream/handle/10183/2702/000375416.pdf?sequence=1) último acesso em 18/04/2021;

"Okapi BM25: a non-binary model", 2008 Cambridge University Press (https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html) último acesso em 18/04/2021;

"Cosine Similarity – Understanding the math and how it works (with python codes)", Selva Prabhakaran (https://www.machinelearningplus.com/nlp/cosine-similarity/) último acesso em 18/04/2021;