UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Identificação de Padrões em Dados de Saúde de Gestantes

Luiz Augusto Matos Tedesco

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Luiz Augusto Matos Tedesco

Identificação de Padrões em Dados de Saúde de Gestantes

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Dilvan Moreira

Versão original

São Paulo

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Tedesco, Luiz
T256i Identificação de Padrões em Dados de Saúde de Gestantes / Luiz Tedesco; orientador Dilvan Moreira. -- São Carlos, 2022.
28 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Saúde. 2. Gestantes. 3. Comorbidades. I. Moreira, Dilvan , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Luiz Augusto Matos Tedesco

Identification of Patterns in Pregnant Women's Health Data on the São Paulo public health system

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Original version

São Paulo

2022

RESUMO

Tedesco, L.A **Identificação de Padrões em dados de Saúde de Gestantes na rede pública de saúde de São Paulo**. 2022. 31p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2022.

Temas como Inteligência Artificial (IA), Big Data, Data Science tem crescido em diversas áreas do conhecimento, ganhando mais e mais espaço no dia a dia, inclusive, na saúde. Em 2020, segundo Pacheco *et al.* (2020), haviam 15.415 registros com "Big Data" na plataforma "Web Of Science" e 326, com "Big Data" e "Healthcare" ou "Health care". Inteligência Artificial é uma área ampla, dentre os trabalhos acima enumerados, existem diversos escopos, inúmeros focos, abordagens e objetivos. O foco deste trabalho será em criar um modelo preditivo, utilizando análise exploratória e grafos, onde seja possível identificar uma tendência de desfecho para novas gestantes com base na série histórica.

Este trabalho usará como base o repositório de dados de saúde do município de São Paulo, criado pelo projeto e-saúdeSP, no escopo do projeto de requalificação das redes de assistência do projeto Avança Saúde, financiado pelo Banco Interamericano de Desenvolvimento (BID).

A partir do conjunto de informações trabalhadas, foi possível criar uma indicação de similaridade entre gestantes que começam o acompanhamento e a série histórica, provendo assim "alertas" ao profissional assistente.

Palavras-chave: Big Data. Inteligência Artificial. Gravidez.

ABSTRACT

Tedesco, L.A **Identification of Patterns in Pregnant Women's Health Data**. 2022. 31p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2022.

Topics such as Artificial Intelligence (AI), Big Data, Data Science have grown in importance in several areas of knowledge, gaining more and more space in everyday life, including health. In 2020, according to Pacheco *et al.* (2020), there were 15,415 records with "Big Data" on the "Web Of Science" platform and 326, with "Big Data" e "Healthcare" or "Health care".
Artificial Intelligence is a broad area, among the works listed above, there are several scopes, numerous focuses, approaches, and objectives. The focus of this work is to create a predictive model, using exploratory analysis and graphs, that can identify outcome trends for newly pregnant women based on a historical series.

This work will be based on the health data repository of the municipality of São Paulo, created by the e-saúdeSP project, in the scope of the Avança Saúde project, funded by the Inter-American Development Bank (IADB).

From the worked data set, it was possible to create an indication of similarity between pregnant women who start to follow up and the historical series, thus proving "alerts" to the professional assistant.

Keywords: Big Data. Artificial Intelligence. Pregnancy.

SUMÁRIO

1 INTRODUÇÃO

O campo da inteligência artificial (IA) estuda sistemas ou máquinas que imitam a inteligência humana para realizar tarefas complexas. Na literatura, existem 3 categorias de IA, aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço, e dentro destas, outras diversas abordagens (algoritmos) sobre suas respectivas implementações. A ascensão da IA levou ao subsequente desenvolvimento de redes neurais artificiais (RNA). Elas são organizadas de forma semelhante aos neurônios do cérebro, com seus múltiplos nós neurais. Os nós de uma rede neural são conectados e enviam dados uns para os outros, para chegar a resposta mais provável. Eles formam um sistema matemático confiável que pode interpretar dados multifatoriais. Fazer essas conexões múltiplas permite que as redes neurais imitem as funções cognitivas do cérebro, como o processo de raciocínio, para identificar a melhor resposta.

Softwares com algoritmos complexos de IA são usados na medicina para analisar grandes quantidades de dados. Essa análise pode auxiliar na prevenção de doenças, diagnóstico e monitoramento de pacientes. Segundo Iftikhar *et al.* (2020), nos últimos anos houve um enorme aumento no crescimento global da inteligência artificial nos sistemas de saúde. De acordo com estatísticas, espera-se que os gastos com IA aumentem de U$2,1 bilhões para U$36,1 bilhões até 2025. A inteligência artificial geralmente requer treinamento e colaboração entre parceiros para se tornar um sucesso em assistência médica. Um exemplo é o sistema Watson for Oncology da IBM (CORP, 2020). Ele fornece opções de tratamento para o câncer, classificadas e baseadas em evidências, para consideração dos médicos responsáveis pelo tratamento, via extração de informações de vários bancos de dados de medicamentos e revistas médicas.

Um dos grandes desafios na análise de dados na área da saúde é a escassez de dados informatizados e estruturados. Com a revolução digital, a expansão da conexão a internet e a expansão do uso dos Prontuários Eletrônico do Paciente (PEPs), dados informatizados e estruturados sobre pacientes passaram a ser armazenados. Desta situação, nasceu um novo problema para a estrutura de saúde pública do Brasil. Cada PEP registra e armazena os dados em seu sistema e, como existem vários sistemas, a informação de saúde do paciente, o seu prontuário de saúde, não fica completo. Visando mitigar essa situação, o ministério da saúde instituiu a Rede Nacional de Dados em Saúde (RNDS). Ela é o projeto estruturante do Conecte SUS, instituído pela portaria GM/MS n. 1.434, de 28 de maio de 2020. Essa rede visa a criação de uma plataforma nacional de interoperabilidade de dados em saúde, com o objetivo de promover a troca de informações entre os pontos da Rede de Atenção à Saúde. Ela permite a transição e continuidade do cuidado ao paciente nos setores público e privado, independente da localização física.

O objetivo da RNDS é a interoperabilidade dos dados para criar uma visão unificada dos PEPs. Essa solução cria um grande repositório de dados clínicos, onde, passa a ser possível se fazer uso em larga escala de análise da dados. A RNDS passa então a criar um grande repositório de informações clínicas abrangendo as mais diversas áreas (RNDS, 2020). A base da interoperabilidade do projeto da RNDS é o Conjunto Mínimo de Dados (CMD) (SNS, 2016).

Visando a aderência ao projeto da RNDS, prevista para 2028, o Município de São Paulo deu início, em 2020, ao projeto e-saúdeSP, (SMS-SP, 2021) via financiamento do BID (SMS-SP/BID, 2021), para a criação de seu próprio repositório de dados clínicos, uma plataforma de telemedicina e um portal para acesso aos dados de saúde do cidadão Paulistano. O e-saúdeSP, Plataforma da Saúde Paulistana, tem como objetivo integrar os prontuários usados na rede de saúde do município, empoderar os usuários do SUS, trazer acesso a informação e viabilizar, por meio da tecnologia da informação, novas modalidades assistenciais aos profissionais de saúde.

1.1 Justificativa e Motivação

Dentro do contexto da saúde, o diagnóstico precoce pode reduzir drasticamente o custo associado ao tratamento, aumentar as chances de sobrevida dos pacientes (mortalidade) e melhorar sua qualidade de vida (morbidade) (NEDEL; ROCHA; PEREIRA, 1999; OPAS, 2017; HEALTH-DATA, 2013).

Doenças crônicas não transmissíveis, como câncer, hipertensão e diabetes, assim como doenças infecto contagiosas, como HIV, sífilis, hepatite e tuberculose, são exemplos de doenças onde o diagnóstico precoce influência no tratamento.

“No campo da medicina, os avanços nos últimos anos levaram a um aumento dramático no volume e na complexidade dos dados biomédicos gerados a partir de indivíduos, experimentos biológicos, hospitais e fatores ambientais” (ANDREU-PEREZ *et al.*, 2015). Isso trouxe novas oportunidades e desafios nas atividades clínicas. O aumento explosivo de dados biomédicos disponíveis excedeu a capacidade dos médicos de extrair todos os dados significativos para obter *insights* sobre doenças complexas, usando métodos estatísticos convencionais. Isso exige um método de análise de nível superior para ajudar os médicos a analisar os dados de forma eficaz (ANDREU-PEREZ *et al.*, 2015; DEO, 2015; OBERMEYER; EMANUEL, 2016). A IA pode aprender as relações potenciais em uma grande quantidade de dados biológicos usando algoritmos complexos. Ela pode usar essas relações para obter *insights* para auxiliar em atividades clínicas"(WANG *et al.*, 2019).

Para a execução deste trabalho foram utilizados dados do "programa mãe paulistana"da secretaria municipal de saúde de São Paulo, esses dados foram escolhidos devido a 3 fatores principais: dados estruturados, volumetria de dados, familiaridade do pesquisador

com o programa, seus indicadores e fluxos;

1.2 Questão de Pesquisa e Objetivos

Considerando a existência do repositório municipal de saúde do projeto e-saúdeSP, a principal questão de pesquisa deste trabalho é:

Com a clusterização dos dados de saúde, provenientes

- do acompanhamento gestacional das mulheres inclusas no "programa mãe paulistana"

, é possível se identificar padrões regionais, assistenciais ou sociais, que influenciem no resultado da gestação?

Diante desta questão de pesquisa, neste projeto são aplicadas técnicas de análise de dados, "K-Clustering", "Decision Tree", "Naive Bayes" e redes neurais artificiais, visando a identificação de padrões e fatores de risco à saúde das gestantes do município.

O objetivo principal é encontrar pontos focais onde seja possível que alguma ação mitigue a taxa de mortalidade e de morbidade materno-infantil. Diante disso, foram definidos os seguintes objetivos para o desenvolvimento deste trabalho:

1. Mapear algoritmos disponíveis na literatura, analisando seus desempenhos na identificação de padrões;
2. Aplicar algoritmos que se destacam na literatura;
3. Inferir pontos de inflexão que possam alterar o resultado da gestação.

2 FUNDAMENTAÇÃO

No “Estudo Bibliométrico” de Pacheco *et al.* (2020) é possível notar a importância, o potencial e as dificuldades da análise de dados na área da saúde. Também é notável o leque de possibilidades, as Figuras 1 e 2 explicitam o crescimento de publicações na área e os termos mais presentes nestas.

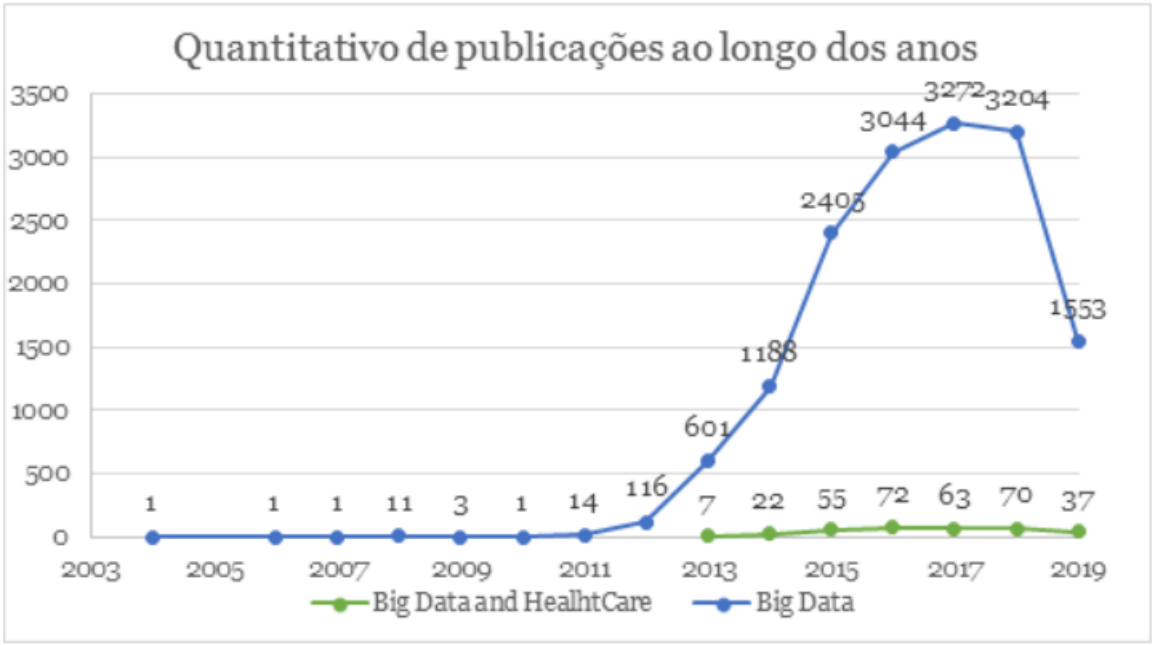


Figura 1 – Quantidade de Publicações

Aprofundando nas citações de Pacheco *et al.* (2020):

O trabalho de Chen, Chiang and Storey (2012) foca na personalização da assistência à saúde, considerando os custos cada vez maiores, fator esse que impede o acesso a tratamentos de saúde para muitos.Enquanto Bates *et al.* (2014) aborda as possibilidades de uso de Big Data na redução de custos na saúde nos Estados Unidos.

Roski, Bo-Linn and Andrews (2014) fala sobre o potencial de se criar valor e suas implicações políticas e éticas; Belle *et al.* (2015) fala dos desafios para aplicar Big Data a saúde. Groves *et al.* (2016) se preocupa com o conceito de um “conjunto mínimo de dados” para protocolos de ensaios clínicos. Murdoch and Detsky (2013) expõe sobre a inevitabilidade do uso de Big Data para contribuir com a qualidade e eficiência dos serviços prestados. Raghupathi and Raghupathi (2014) fala do potencial e desafios a serem superados.

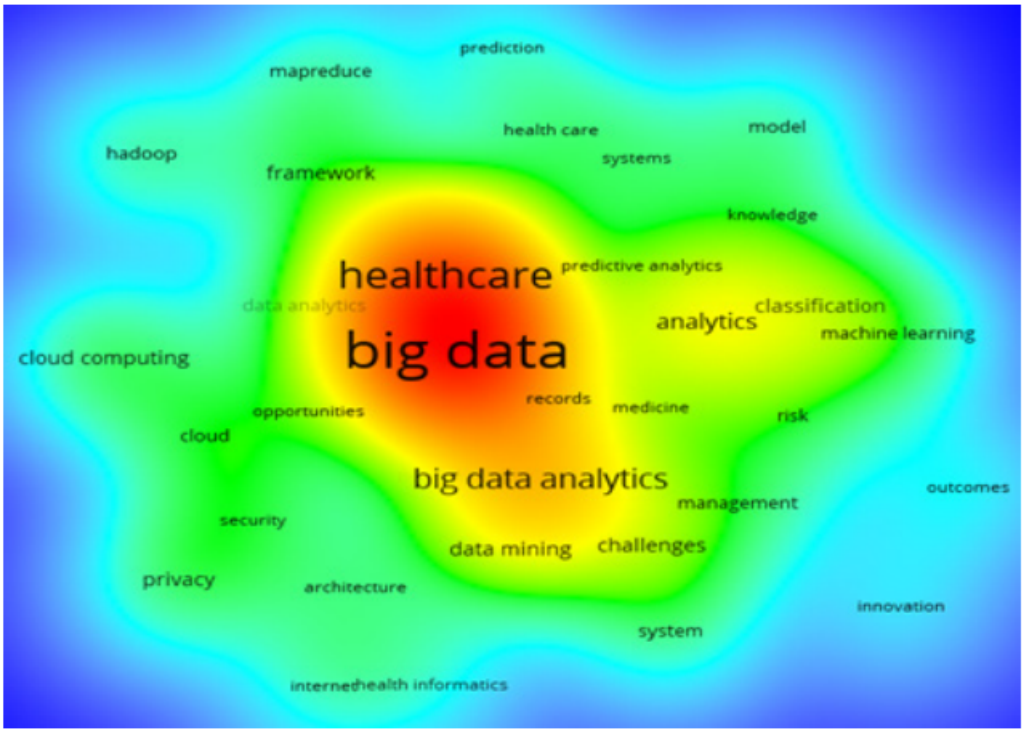


Figura 2 – Mapa de Calor

Bates *et al.* (2014) e Beam and Kohane (2018) decorrem sobre a análise preditiva como forma de identificar e preencher lacunas no sistema de saúde. Bates *et al.* (2014) levanta a questão do consumo de dados de múltiplas fontes, as questões de privacidade, ética e a necessidade da adoção de padrões dos dados; Beam and Kohane (2018) faz associações entre Machine Learning e Big Data, e sobre o quanto destes são influenciados e decididos por humanos.

Aumentando a resolução da pesquisa, saindo de *Health Care* e indo para "gestação". Em específico, encontrou-se artigos referentes ao tratamento de infertilidade (WANG *et al.*, 2019), ao uso de drogas (medicamentos) durante a gestação (DAVIDSON; BOLAND, 2020), sobre associações de padrões de comportamento de saúde, saúde mental e saúde autoavaliada (OFTEDAL *et al.*, 2019), sobre o impacto fisiológico da "gestação" na saúde da mulher e em suas futuras gestações (STUDNICKI *et al.*, 2020).

Importante ressaltar que a maioria dos trabalhos acima citados tratam de teoria, questões éticas e implicações, nenhum dos trabalhos citados possui um escopo na área deste trabalho,

O foco deste trabalho foi: Criar um modelo preditivo, utilizando análise exploratória e grafos, onde seja possível identificar uma tendência de desfecho para novas gestantes com base na série histórica. Para tal ficam estabelecidos dois objetivos: 1 - Identificar quais os dados mais relevantes dentro do conjunto de dados completo, para alcançar o objetivo

2; 2 - Criar um modelo preditivo, que indique a tendência do desfecho da gestação em novos casos. Para tal, serão utilizadas técnicas de identificação de correlação para atingir o objetivo 1, e análise de proximidade em grafos simples para atingir o objetivo 2.

2.1 Teoria

A teoria dos grafos é um arcabouço natural para o tratamento matemático de redes. As medidas utilizadas em redes complexas são formalizadas matematicamente a partir de um grafo.(CABRAL, 2013). É um ramo de conhecimento bastante antigo, pertercente à matemática que estuda as relações entre os objetos de um determinado conjunto. Foi introduzida no século XVIII pelo matemático suiço Leonhard Euler, que utilizou grafos para resolver o problema que conhecemos como As sete pontes de Königsberg. Grafos possuem aplicações nas mais diferentes áreas e é possível desenvolver técnicas de visualização que auxiliam os usuários a entender visualmente os dados trabalhados. Existem muitas maneiras de se representar grafos, cada uma com suas vantagens e desvantagens, entre elas: grafos não direcionados, direcionados, ponderados ou não.

De forma informal, grafos podem ser definidos como uma abstração que permite codificar relacionamentos entre pares de objetos, onde os relacionamentos e os objetos variam de acordo com a área de aplicação, podendo ser por exemplo: possíveis amizades em uma rede social. A definição formal é:

$$G = (V(G), E(G), \Psi G) \tag{2.1}$$

Onde V(G) é o conjunto não vazio de vértices, E(G) são as arestas e Ψ G a função que associa cada aresta de G a um par de vértices de G.

Quando se usa grafos em problemas que a IA busca trabalhar, faz-se necessário percorrê-los.

Deve-se ter uma forma sistemática de visitar as arestas e os vértices desses grafos. Existem problemas a serem tratados, como a eficiência, onde não devem haver repetições (desnecessárias) de visitas a um vértice e/ou aresta, ou a completude, afinal todos os vértices e/ou arestas devem ser visitados.

Há duas estratégias básicas para pesquisar/percorrer/caminhar sobre um grafo: a busca em largura (BreadthFirst Search, BFS) e a busca em profundidade (Depth-First Search, DFS) ambas possuem vantagens e desvantagens. Em 2016, foi proposta uma nova abordagem ao BFS e FDS, o *node2vec: Scalable Feature Learning for Networks* que será usado neste trabalho. (NODE2VEC. . . , 2016)

O node2vec é uma estrutura algorítmica para a aprendizagem representacional em grafos. Dado qualquer grafo, o node2vec pode aprender representações de características

contínuas para os nós, que podem então ser usadas para várias tarefas de aprendizagem de máquina. O algoritmo propõe um mapeamento dos nós para um espaço de baixa dimensão de características que maximizam a probabilidade de preservar a rede de nós vizinhos, através da introdução de uma *random walk* parametrizável.

BFS e DFS representam cenários extremos em sua execução e, por consequência, cada um trás implicações específicas no conteúdo aprendido. As tarefas de aprendizado costumam trabalhar dois (2) tipos de similaridades: homofilia e equivalência estrutural. De forma resumida, homofilia, itens similares enquanto que na equivalência estrutural refere-se ao "papel"do item na estrutura dos dados.

Vizinhanças “criadas” via BFS correspondem fortemente com a equivalência estrutural enquanto aquelas “criadas”, via DFS, agregam similaridade por homofilia. Entretanto, no mundo real, os dados costumeiramente têm estes 2 aspectos ao mesmo tempo. O node2vec faz uso de *second order random walk* para “interpolar” entre as 2 abordagens.

O node2vec consome um grafo construído ao redor da variável de interesse. Devido a isso, neste trabalho, o conjunto de dados foi “trabalhado” para construir o grafo que servirá de base ao Node2Vec. Entre as adequações, vale citar o método de Tukey para remoção de outliers, consolidação de variáveis em dados que sejam compatíveis com os indicadores da SMS-SP (SMS-SP, 2022) e codificação das variáveis numéricas e categóricas de forma a serem trabalhadas em grafos.

2.2 Modelo de Dados DataLake e-saúdeSP

A seguir estão descritos os modelos de dados disponíveis na plataforma de análise de dados do projeto e-saúdeSP.

2.2.1 Agendamento

- Especialidade
- Procedimento
- Tipo de Consulta
- Data da Consulta

2.2.2 Procedimento

- Classificação Brasileira de Ocupações (CBO) do solicitante
- Data do procedimento

2.2.3 Programa Mãe Paulistana

2.2.3.1 Acolhimento

- Data de acolhimento
- Data da Ultima menstruação
- Data da Prevista do parto
- Unidade de Pré-Natal

- Quantidade de Gestações Anteriores
- Quantidade de Filhos nascidos vivos
- Quantidade de Filhos natimortos
- Quantidade de Partos vaginais
- Quantidade de Partos cesarea
- Quantidade de Abertos
- Quantidade de Tempo Decorrido (semanas)
- Situação Vacinal no acolhimento

2.2.3.2 Atendimento

- Idade Gestacional Corrigida
- Idade Gestacional Calculada
- Pressão arterial Mínima
- Pressão arterial Máxima
- Altura Uterina
- Peso
- Altura
- IMC
- Situação Vacinal no atendimento
- Classe Profissional
- Risco Gestacional no Atendimento
- Classificação Risco no Atendimento
- Encaminhamento Risco?
- Encaminhamento para qual unidade?
- Situação Vacinal no Acolhimento
- Quantidade de Partos Cesarea
- Quantidade de Abertos
- Quantidade de Tempo Decorrido (semanas)
- Situação Vacinal no Acolhimento
- Quantidade de Partos cesarea
- Quantidade de Abertos
- Quantidade de Tempo Secorrido (semanas)
- Situação Vacinal no Acolhimento
- Batimento Cardiaco Fetal

2.2.3.3 Deslocamento

- Precisa de Auxilio Transporte Público?
- Quantidade de Deslocamento
- Data da Solicitação
- Data da Entrega

2.2.3.4 Exames

- Exame Solicitado
- Resultados

2.2.3.5 Risco

- Condição Prévia de Saúde

2.2.3.6 Laudo Exame

- Unidade Executora
- Tipo Atendimento
- Modalidade
- Exame
- Data
- Procedimento

Os dados acima citados foram captados do SIGA (Sistema Integrado de Gestão de Assistência) SAÚDE, Sistema Matrix (Sistema de Captação/tramitação de exames laboratoriais), FIDI (Fundação Instituto de Pesquisa e Estudo de Diagnóstico por Imagem) e de prontuários das organizações sociais que atuam no município. A base de dados é alimentada diariamente, via integrações. A Tabela 2 mostra o volume de atendimentos no dia 28/01/2022.

Tabela 1 – Volumetria dos dados

SIGA - Sistema Integrado de Gestão de Atendimento	
Atendimentos	157.361.943
Agendamentos	1.402.275
PEP´s	
Atendimentos	17.664.462
Problema	6.947.302
Procedimento	12.366.310
Mãe Paulistana	
Acompanhamentos	2.288.978
Atendimento	7.309.971
Batimentos Cardíacos Fetais	5.718.327
Deslocamento	658.864
Exames	12.827.898
Gestante de Risco	646.980
Exames e Vacinas	
Pedido de Exames	4.062.573
Resultado de Exames	270.156
Resultado de Exames Procedimento	2.022.084
Resultado de Exames Valorado	14.030.122
Laudos	1.743.848
Vacinas	13.585.522

3 METODOLOGIA

3.1 Autorização para uso

O passo base para a estruturação do projeto foi identificar a forma de obter permissão de uso e acesso ao conjunto de dados do projeto e-saúdeSP junto a secretaria municipal de saúde de São Paulo (SMS) (CEP-SMS/SP, 2021). Após a tramitação via plataforma Brasil e de posse da autorização e do acesso aos dados, foi necessário compreender a relação entre as tabelas, chaves primárias e estrangeiras.

3.2 Coleta de Dados

A estrutura de dados do projeto e-saúdeSP estava armazenada num servidor SQL, (Microsoft SQL Azure). Para acesso a ela, foi utilizado o driver SQL para Python, Pyodbc (MICROSOFT, 2022). Os dados inicialmente selecionados para o estudo estavam dispostos em 11 tabelas conforme mostrado na Tabela 2 e a Figura 3:

Tabela 2 – Distribuição dos dados

Tabelas BD e conteúdo resumido	
Gestação	Estrutura de Dados Central da Gestação
Pessoa Física	Atributos pessoais da gestantes
Risco	Fatores de Risco identificados
Deslocamento	Transporte para acompanhamento
Atendimento	Ficha de atendimento de pré-natal
Batimento Cardiaco Fetal	BCF do atendimento
Exames Gestação	Exames solicitados/registrados no atendimento de pré-natal
Exames Pessoa Física	Exames não relacionados a gestação
Procedimento Exames	Precedimentos destes exames
Resultados exames	Resultados
Laudo Exame	Laudos

Devido a grande quantidade de dados e limitações computacionais, foi necessário serializar o processo de extração dos dados, necessários para o estudo das variáveis, e criação do experimento. A partir do primeiro conjunto de dados analisados, concluiu-se que o conjunto de interesse poderia ser aprimorado, reduzindo o tempo de extração de dados. Nesta etapa, temos a primeira "transformação" deste trabalho. A variável de interesse "desfecho da gestação" existe no banco como um conjunto de informações de 2 variáveis: "Data de parto" e "Interrupção motivo". A partir desta primeira conclusão, a query foi atualizada para trazer apenas as variáveis de interesse.

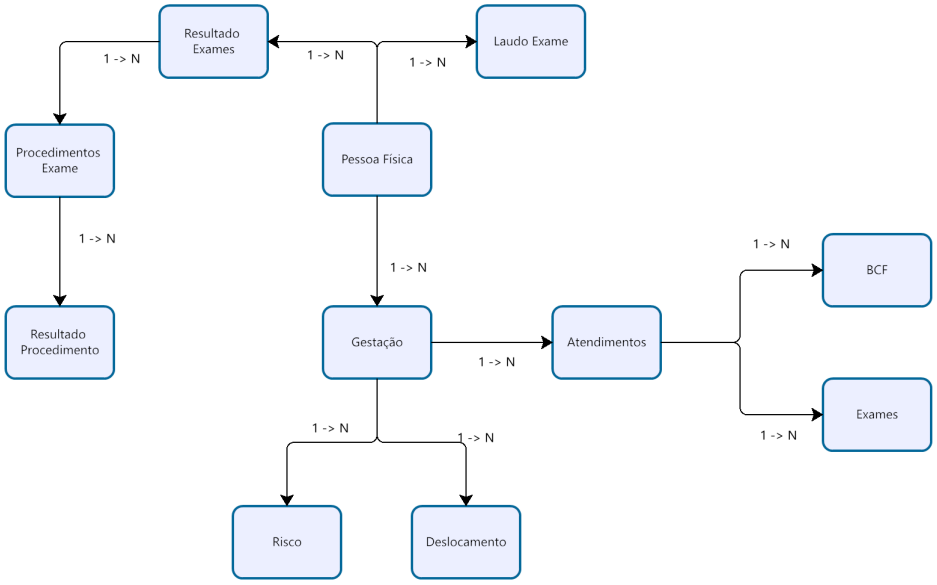


Figura 3 – Diagrama relacional

3.3 Seleção do Conjunto de Dados a utilizar

Solucionado a etapa de extração de dados e a higienização da base, foram realizadas conversas com médicos da SMS SP, onde escolheu-se um subconjunto de 141 variáveis e indicadores. O primeiro passo foi a remoção de variáveis cujo preenchimento não fosse significativo no montante total. Uma característica percebida na análise dos dados foi a necessidade de combinações e/ou comparações para que os mesmos pudessem passar a representar uma informação. Nesta linha, foi calculada a quantidade de consultas que a gestante teve, a idade gestacional aproximada no acolhimento no serviço de saúde, sua idade neste acolhimento, PA (Pressão Arterial) mínima e máxima, a variação do IMC(Índice de massa corporal) da gestante e a quantidade de exames realizados durante a gravidez, por tipo de exame.

Devido a complexidade para se relacionar os dados das 11 tabelas iniciais, após testes e deliberações, escolheu-se remover para este experimento os dados de:

“Laudo de exames”, por ser necessário uma análise textual (Processamento de linguagem natural) para entender seus resultados.

“Resultado exames”, por ser necessário aprofundamento nos inúmeros analitos derivados de cada exame e definição de padrões e métodos de comparação que são específicos para cada analito.

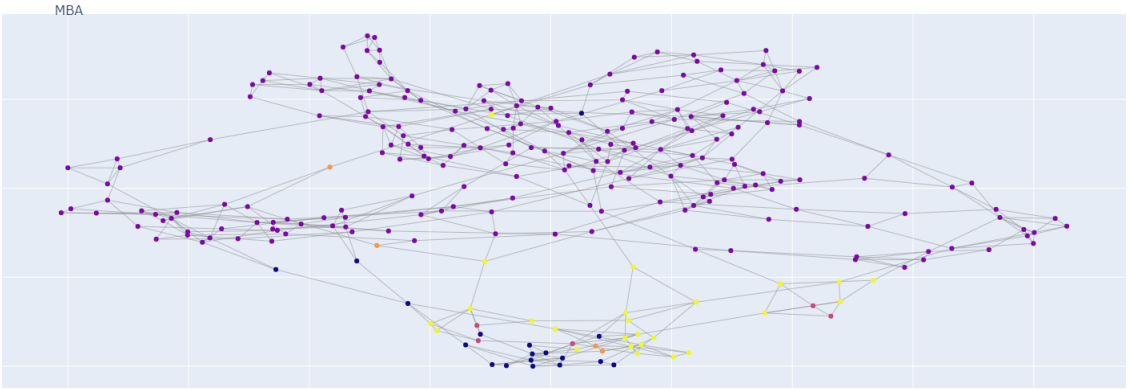
Os dados restantes foram separados em 2 tipos de atributos: Categórico e Numérico: Início do acompanhamento(precoce, normal, tardio) atributo categórico, quantidade de exames de sangue, atributo numérico.

Categórico:	‘QNT_CONSULTAS’, ‘IG_ACOLHIMENTO’, ‘DESFECHO’,‘CD_RACA’, ‘BL_RISCO_GESTACIONAL_ATENDIMENTO’, ‘BL_PRE_NATAL_ESTABELECIMENTO_ACOLHIMENTO’
Númérico:	‘QT_GESTACOES_ANTERIORES’, ‘QT_FILHOS_NASCIDOS_VIVOS’, ‘QT_FILHOS_NATIMORTOS’, ‘QT_PARTOS_VAGINAL’, ‘QT_PARTOS_CESAREA’, ‘QT_ABORTOS’,‘NR_TEMPO_INTERNACAO’, ‘VL_PESO_BEBE’, ‘VL_PRESSAO_MINIMA’, ‘VL_PRESSAO_MAXIMA’, ‘VL_ALTURA_UTERINA’, ‘VL_PESO’, ‘VL_ALTURA’, ‘VL_IMC’

Para alguns atributos categóricos foi necessário complementar os dados pois algumas informações estavam implícitas na sua ausência de preenchimento e para facilitar a análise posterior. Tais itens foram traduzidos para uma categoria visando facilitar a leitura e interpretação. Os atributos numéricos passaram por uma etapa de remoção de outliers através do método de Tukey. Para representação dos dados acima citados, foi utilizado o OneHotEncoder para os categóricos e StandardScaler para os numéricos, ambos do pacote scikit-learn. O OneHotEncoder cria uma coluna binária para cada categoria e retorna uma matriz esparsa ou matriz densa (ONEHOTENCODER,) enquanto que o StandardScaler padroniza as categorias removendo a média e escala a variância a uma unidade(STANDARDSCALER,).

Feitas as devidas transformações, foi criado um grafo para alimentar o algorítimo Node2Vec responsável por buscar/identificar relações.

A figura 3.3 mostra o grafo criado na etapa preliminar.



4 ANÁLISE DE RESULTADOS E CONCLUSÃO

Na proposição deste trabalho, foram elencados os seguintes objetivos:

1. Mapear algoritmos disponíveis na literatura, analisando seus desempenhos na identificação de padrões
2. Aplicar algoritmos que se destacam na literatura
3. Inferir pontos de inflexão que possam alterar o resultado da gestação

O algorítimo escolhido para a execução do projeto foi o node2vec devido a sua capacidade de trabalhar de forma flexível, como apontado na teoria deste, e a avaliação de que o mesmo é 12.6 % mais efetivo do que o estado da arte (NODE2VEC. . . , 2016). O resultado obtido com o sub-conjunto utilizado foi inconclusivo, utilizou-se um subconjunto de dados devido a limitações de memória RAM, também não foi possível identificar pontos de inflexão que pudessem alterar o resultado da gestação, porém é possível indicar quais gestantes são mais similares ao caso em análise, desta forma, provendo ao profissionais assistente uma visão de casos "análogos", com o intuito de substanciar a "direção" na qual aquela gestação está seguindo.

As ações listadas a seguir poderão ser conduzidas para melhorar o projeto:

1. Serializar a execução do processo desenvolvido neste projeto para que seja possível executá-lo com a completude de dados disponíveis;
2. Analisar as variáveis não utilizadas na etapa atual e identificar como inseri-las no processo;
3. Analisar outras variáveis não cogitadas anteriormente e identificar como inseri-las no processo, tais análises podem ser feitas usando algoritmos de seleção de atributos, verificando quais atributos são mais importantes;
4. Incluir uma relação temporal entre algumas variáveis, como por exemplo: Data das Consultas;
5. Incluir a estrutura regionalizada da rede de atenção a saúde da SMS SP;
6. Criar um mecanismo para verificar a acurácia da proposição.

Para um futuro trabalho é possível agregar outros dados e conceitos como o de regionalização dos serviços de saúde, assim como a existência de serviços especializados,

alta complexidade. Também é possível incluir informações sobre notificações de agravos de saúde pública e atendimentos em unidades de urgência e emergência.

Assim como exposto no início deste trabalho o diagnóstico precoce pode reduzir drasticamente o custo associado ao tratamento, aumentar as chances de sobrevida dos pacientes (mortalidade) e melhorar sua qualidade de vida (morbidade) (NEDEL; ROCHA; PEREIRA, 1999; OPAS, 2017; HEALTH-DATA, 2013).

A Figura 4 mostra o resultado da busca de similaridade, que poderia ser utilizada como apoio a tomada de decisão pelo profissional assistente, visando um "diagnóstico precoce" e deste forma, melhorando morbidade e mortalidade. Considerando a enorme complexidade do ciclo gestacional, envolvendo desde fatores biológicos a socioeconômicos, existe uma unidade singular a cada paciente, e, este "alerta" de proximidade a gestantes com desfechos indesejados pode auxiliar o profissional a dedicar um tempo para uma análise mais cuidadosa da imensidade de fatores envolvidos.

```
Name: 28495, dtype: object
Gestantes mais similares:
Gestante  ('239', 0.946516752243042) ID_GESTANTE= 26568  similaridade= 0.946516752243042
Gestante  ('72', 0.8970493078231812) ID_GESTANTE= 7585  similaridade= 0.8970493078231812
Gestante  ('73', 0.8863465785980225) ID_GESTANTE= 7619  similaridade= 0.8863465785980225
Gestante  ('130', 0.8718999624252319) ID_GESTANTE= 14279  similaridade= 0.8718999624252319
Gestante  ('96', 0.859484851360321) ID_GESTANTE= 9690  similaridade= 0.859484851360321
Gestante  ('199', 0.8373269438743591) ID_GESTANTE= 22278  similaridade= 0.8373269438743591
Gestante  ('225', 0.8259339928627014) ID_GESTANTE= 24279  similaridade= 0.8259339928627014
Gestante  ('197', 0.8143825531005859) ID_GESTANTE= 21840  similaridade= 0.8143825531005859
Gestante  ('31', 0.8134176135063171) ID_GESTANTE= 3809  similaridade= 0.8134176135063171
Gestante  ('213', 0.7973641753196716) ID_GESTANTE= 23171  similaridade= 0.7973641753196716
```

Figura 4 – Busca de similaridade

REFERÊNCIAS

ANDREU-PEREZ, J. *et al.* Big data for health. **IEEE journal of biomedical and health informatics**, IEEE, v. 19, n. 4, p. 1193–1208, 2015.

BATES, D. W. *et al.* Big data in health care: using analytics to identify and manage high-risk and high-cost patients. **Health affairs**, v. 33, n. 7, p. 1123–1131, 2014.

BEAM, A. L.; KOHANE, I. S. Big data and machine learning in health care. **Jama**, American Medical Association, v. 319, n. 13, p. 1317–1318, 2018.

BELLE, A. *et al.* **Big Data Analytics in Healthcare, BioMed research international. Volume 2015, Article ID 370194**. [*S.l.: s.n.*]: Hindwai Publishing Corporation, 2015.

CABRAL, R. da S. Estudo da variabilidade de medidas em redes complexas. Universidade Federal de Minas Gerais, 2013.

CEP-SMS/SP, C. de Ética em Pesquisa da Secretaria Municipal de S. **Autorização SMS - CEP**. 2021. Available at: https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/comite_de_etica/index.php?p=283229.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. **MIS quarterly**, JSTOR, p. 1165–1188, 2012.

CORP, I. **IBM Watson Health in Oncology**. 2020. Acessado em (...). Available at: https://www.ibm.com/downloads/cas/NPDPLDEZ.

DAVIDSON, L.; BOLAND, M. R. Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence. **Journal of pharmacokinetics and pharmacodynamics**, Springer, p. 1–14, 2020.

DEO, R. C. Machine learning in medicine. **Circulation**, Am Heart Assoc, v. 132, n. 20, p. 1920–1930, 2015.

GROVES, P. *et al.* The'big data'revolution in healthcare: Accelerating value and innovation. Center for US Health System Reform Business Technology Office, 2016.

HEALTH-DATA. **CARGA DE DOENÇA GLOBAL: GERANDO EVIDÊNCIAS, POLÍTICA DE ORIENTAÇÃO**. 2013. Available at: http://www.healthdata.org/sites/default/files/files/policy_report/2013/GBD_GeneratingEvidence/IHME_GBD_GeneratingEvidence_FullReport_PORTUGUESE.pdf.

IFTIKHAR, P. *et al.* Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. **Cureus**, Cureus Inc., v. 12, n. 2, 2020.

MICROSOFT. **Driver SQL Python**. 2022. Available at: https://docs.microsoft.com/pt-br/sql/connect/python/pyodbc/python-sql-driver-pyodbc?view=sql-server-ver16.

MURDOCH, T. B.; DETSKY, A. S. The inevitable application of big data to health care. **Jama**, American Medical Association, v. 309, n. 13, p. 1351–1352, 2013.

NEDEL, F. B.; ROCHA, M.; PEREIRA, J. Anos de vida perdidos por mortalidade: um dos componentes da carga de doenças. **Revista de Saúde Pública**, SciELO Brasil, v. 33, p. 461–469, 1999.

NODE2VEC: Scalable Feature Learning for Networks. 2016. Available at: https://arxiv.org/pdf/1607.00653v1.pdf.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future—big data, machine learning, and clinical medicine. **The New England journal of medicine**, NIH Public Access, v. 375, n. 13, p. 1216, 2016.

OFTEDAL, S. *et al.* Associations of health-behavior patterns, mental health and self-rated health. **Preventive medicine**, Elsevier, v. 118, p. 295–303, 2019.

ONEHOTENCODER. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. Accessed: 2022.01.15.

OPAS. **Diagnóstico precoce do câncer salva vidas e reduz custos de tratamento**. 2017. Available at: https://www.paho.org/pt/noticias/3-2-2017-diagnostico-precoce-do-cancer-salva-vidas-e-reduz-custos-tratamento.

PACHECO, R. R. *et al.* Big data em healthcare–um estudo bibliométrico. **Revista Ibérica de Sistemas e Tecnologias de Informação**, Associação Ibérica de Sistemas e Tecnologias de Informacao, n. E28, p. 739–751, 2020.

RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. **Health information science and systems**, Springer, v. 2, n. 1, p. 1–10, 2014.

RNDS. **Rede Nacional de Dados em Saúde**. 2020. https://www.gov.br/saude/pt-br/assuntos/rnds. Accessed: 2021-09-30.

ROSKI, J.; BO-LINN, G. W.; ANDREWS, T. A. Creating value in health care through big data: opportunities and policy implications. **Health affairs**, v. 33, n. 7, p. 1115–1122, 2014.

SMS-SP. **Rede Mãe Paulistana - Programa da Secretaria Municipal da Saúde de São Paulo**. 2022. Acessado em (...). Available at: https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/programas/mae_paulistana/index.php?p=5657.

SMS-SP, S. M. de Saúde de S. P. **e-saúdeSP, Plataforma de Saúde Paulistana**. 2021. https://www.municipal.com.br/aplicativo-e-saudesp-app-plataforma-da-saude-paulistana/. Accessed: 2021-09-30.

SMS-SP/BID, S. M. de Saúde de S. P. **Avança Saúde SP**. 2021. https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/programas/index.php?p=256177. Accessed: 2021-09-30.

SNS, S. N. d. S. **Conjunto Mínimo de Dados - CMD**. 2016. https://conjuntominimo.saude.gov.br/#/cmd. Accessed: 2021-09-30.

STANDARDSCALER. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed: 2022.01.15.

STUDNICKI, J. *et al.* Pregnancy outcome patterns of medicaid-eligible women, 1999-2014: a national prospective longitudinal study. **Health Services Research and Managerial Epidemiology**, SAGE Publications Sage CA: Los Angeles, CA, v. 7, p. 2333392820941348, 2020.

WANG, R. *et al.* Artificial intelligence in reproductive medicine. **Reproduction**, Bioscientifica Ltd, v. 158, n. 4, p. R139–R154, 2019.