Aprendizado de máquina para diagnóstico de Diabetes Mellitus.

Henrique Oliveira Serra[1], Vinício da Silva Nascimento[2],
Orientadora: Prof.ª M.ª Liszeila Reis Abdala Martingo
Coorientador: Prof.º José Alexandre Ducatti

e-mail: henrique.serra@fatec.sp.gov.br; vinicio.nascimento@fatec.sp.gov.br;
Orientadora e-mail: liszeila.martingo@fatec.sp.gov.br;
Coorientador e-mail: jose.ducatti@fatec.sp.gov.br;

Resumo: O uso da tecnologia é cada vez mais frequente, sendo assim, é necessário que os projetos relacionados a área da saúde estejam cada vez mais correlacionados aos de tecnologia. Nesse contexto existem inúmeras vertentes a serem exploradas. O presente trabalho tem por objetivo utilizar modelos de predição para diagnóstico precoce de diabetes mellitus, através do aprendizado de máquina, a análise de dados se dá através de técnicas de regressão logística para treinamento do modelo, que contém informações sobre a doença. Para isso utilizamos diversas tecnologias para identificar um padrão na ocorrência da doença e revelar possíveis casos positivos em seu estágio inicial. De modo a alertar e contribuir com a população para um diagnóstico precoce, possibilitando controle da doença.

Palavras-chave: Aprendizado supervisionado. Saúde. Diabetes Mellitus II. Machine Learning. Diagnostico. Predição.

Abstract: *The use of technology is increasingly frequent, so it is necessary that those related to a health area and projects are increasingly correlated to health projects. In this context, there are numerous aspects to be explored. The work aims to use prediction models for early diagnosis of diabetes mellitus, through machine learning, and an analysis of data that contains information about the disease. For this, we use several technologies to identify a pattern in the occurrence of the disease and possible positive cases in its initial stage. In order to alert and contribute to the population for an early diagnosis, predicting the control of the disease.*

Keywords: Supervised learning. Health. Diabetes Mellitus II. Machine Learning. Diagnosis. Prediction.

1 Introdução

O Brasil é o 5º país em incidência de diabetes mellitus tipo 2 no mundo, com 16,8 milhões de doentes adultos (20 a 79 anos), perdendo apenas para China, Índia, Estados Unidos e Paquistão. A estimativa da incidência da doença em 2030 chega a 21,5 milhões. (BRASIL, 2020).

A crescente urbanização e a mudança de hábitos de vida (por exemplo, maior ingestão de calorias, aumento do consumo de alimentos processados, estilos de vida sedentários) são fatores que contribuem para o aumento da prevalência de diabetes mellitus tipo 2 em nível social. Enquanto a prevalência global de diabetes nas áreas urbanas é de 10,8%, nas áreas rurais é menor, de 7,2%. No entanto, essa lacuna está diminuindo, com a prevalência rural aumentando. (BRASIL, 2020).

Ao passo em que as doenças metabólicas crescem e alcançam cada vez uma maior parcela da população mundial, cresce também a utilização de tecnologias para detecção, prevenção e alerta.

Utilizaremos os índices de diabetes mellitus tipo 2 para elaboração das análises de predição pois trata-se de uma doença metabólica desencadeada através de maus hábitos de vida, tais como sedentarismo, alimentação e obesidade. (TENORIO; PINHEIRO, 2019).

De modo geral, as aplicações na área da saúde são técnicas de aprendizado supervisionado para algoritmos de classificação. Isso significa que, por meio de um banco de dados de vários pacientes contendo a relação de seus sintomas e diagnósticos, busca-se encontrar padrões de sintomas para cada enfermidade. Desta forma, a aplicação é capaz de informar se o paciente tem ou não determinada enfermidade, de acordo com os sintomas que apresenta. (SAHU, 2020).

Temos desse modo a junção ideal para a detecção precoce de Diabetes Mellitus.

2 Justificativa

A relevância desse estudo se faz necessário não só para a aplicação de conhecimentos voltados a tecnologia, mas também a contribuição com a sociedade na detecção preventiva do diagnóstico do Diabetes Mellitus Tipo 2, em que na grande maioria das vezes acaba por ser diagnosticado tardiamente, muitas vezes comprometendo a saúde do indivíduo, e dificultando o tratamento e controle. O uso de tecnologias de predição, possibilitam tratamento precoce e eficaz sem comprometer o indivíduo.

3 Objetivo(s)

Os principais objetivos que podemos destacar com esse estudo são, as aplicações de sistemas de Machine Learning, Inteligência Artificial, para a detecção e rápido controle de doenças, trazendo a público a necessidade de melhorias em diagnósticos, tais como assertividade, agilidade e antecipação.

- Compreender os sinais e sintomas relacionados ao Diabetes Mellitus;
- Entender conceitos de Machine Learning.
- Aplicar técnicas de regressão logística para treinar o modelo de detecção;
- Comparar modelos para encontrar a melhor acurácia possível para o projeto;
- Elaborar interface de usuário para interação e posterior resposta;
- Apresentar e discutir os resultados obtidos e sua utilização.

4 Fundamentação Teórica

Neste capítulo será abordado os principais conceitos relacionados a Machine Learning, inteligência artificial e os principais métodos utilizados para chegar ao resultado.

4.1 Python

Python é uma linguagem de propósito geral de alto nível, multiparadigma, suporta o paradigma orientado a objetos, imperativo, funcional e procedural. Possui tipagem dinâmica e uma de suas principais características é permitir a fácil leitura do código e exigir poucas linhas de código se comparado ao mesmo programa em outras linguagens. Devido às suas características, ela é utilizada, principalmente, para processamento de textos, dados científicos e criação de Computer Graphic Imagery (CGI) para páginas dinâmicas para a web. (WIKIPÉDIA, 2022)

Guido Van Rossum criou o Python em 1989. Ele trabalhava no Centrum Voor Wiskunde en Informatica (CWI) no início dos anos 1980, e seu trabalho era implementar a linguagem de programação conhecida como ABC.

Durante o final dos anos 1980, enquanto ainda estava no CWI, ele começou a procurar por uma linguagem de script que tivesse sintaxe semelhante ao ABC, mas que tivesse acesso às chamadas de sistema do Amoeba. Após procurar e não encontrar nenhuma linguagem que

atendesse às suas necessidades, Rossum decidiu projetar uma linguagem de script simples que pudesse superar as inadequações do ABC. (LIMA, 2021)

O Python se torna especial, pela sua facilidade no aprendizado, ensino entendimento e instalação dos módulos necessários para sua utilização. Gratuito, ele se torna uma ferramenta abrangente a todos. (BRASIL, 2020)

4.2 Pandas

O pandas é uma ferramenta de análise e manipulação de dados de código aberto construída sobre a linguagem de programação Python.

O pandas pode ser utilizado para manipulação de dataset tais como, limpeza de dados, normalização de dados, alterações entre caracteres e números.

É uma ferramenta de fácil utilização, ou seja, com poucos comandos é possível realizar boa parte do trabalho.

Sua grande importância fica a cargo da fase de pré-processamento, onde as bases de dados (dataset), devem ser formatadas de modo a extrair posteriormente o modelo estatístico.

Cada modelo terá por objetivo identificar os padrões obtidos no pré-processamento dos dados, e revelar características não visíveis. Após essa fase, temos a etapa de análise e relevância dos dados obtidos, a fim de validá-los.

A etapa de pré-processamento de dados (segunda etapa) costuma ser a mais trabalhosa em qualquer projeto relacionado à ciência de dados, ocupando tipicamente 80% do tempo consumido. É nesta fase que são realizadas as tarefas de seleção, limpeza e transformação dos dados que serão utilizados pelo algoritmo de Machine Learning / Estatística. O objetivo da seleção de dados é coletar e reunir todos os dados que sejam relevantes para a resolução do problema de ciência de dados definido (por exemplo, combinar dados dos sistemas corporativos da empresa com dados disponibilizados na internet). Limpeza, significa eliminar sujeira e informações irrelevantes. Por fim, transformação consiste em converter os dados de origem para um outro formato, mais adequado para ser usado pelo algoritmo. As atividades de seleção, limpeza e transformação de dados são comumente referenciadas como atividades de Data Wrangling, Data Munging ou Data Preparation. (CORRÊA, 2020).

Temos assim a utilização dessa ferramenta de suma importância e que possui um público alvo independentemente do nível de conhecimento ou experiencia.

4.3 Dataset kaggle

O presente dataset utilizado teve como origem a pesquisa Behavioral Risk Factor Surveillance System (BRFSS), realizada anualmente pela Centro de Controle e Prevenção de Doenças (CDC), que conteve cerca de 441.455 entrevistados e 330 perguntas para obtenção do dataset.

O Kaggle faz parte do grupo de empresas atreladas ao Google, diretamente ligado ao aprendizado de Machine Learning. Possui uma vasta gama de datasets dentro os mais variados tipos, sempre com o intuito de auxiliar no proposito do aprendizado, disponibilizando material suficiente necessário para análise de diferentes situações aplicando machine learning. (RIBEIRO, 2018).

Resumidamente, o Kaggle dentro da sua plataforma pode hospedar competições de Data Science públicas, privadas e acadêmicas. As competições patrocinadas por empresas oferecem prêmios em dinheiro pela melhor solução. Também existem competições de aprendizado (disponibilizadas pelo próprio Kaggle ou por empresas para treinamento de habilidades). A plataforma também armazena e disponibiliza dados sobre assuntos diversos (chamados datasets) e possui fóruns para troca de conhecimentos entre seus usuários (RIBEIRO, 2018).

O Kaggle é uma ótima plataforma para aprendizado e prática de Data Science, abrangendo níveis variados de conhecimento entre seus usuários. A possibilidade de troca de informações é útil tanto para quem está iniciando seu aprendizado nessa área, quanto para os mais avançados. As competições são bastante desafiadoras, mostrando problemas reais de negócio a serem resolvidos. Acredito que vale a pena dar uma olhada e acompanhar os tutoriais para quem está iniciando e depois se arriscar em competições de aprendizado para então testar os conhecimentos adquiridos numa competição real. (RIBEIRO, 2018)

4.4 Scikit-learn

O Scikit-Learn fornece ferramentas importantes para os vários momentos do ciclo de projetos de Machine Learning, tal como MLPClassifier, que é uma rede neural de classificador de perceptron de várias camadas, este modelo utiliza a função log-loss ou gradiente descendente estocástico. A perda logarítmica (relacionada à entropia cruzada) mede

o desempenho de um modelo de classificação em que a entrada de previsão é um valor de probabilidade entre 0 e 1. (LIMA, 2020).

O scikit-learn é a biblioteca mais conhecida dentro do universo Python, sendo de código aberto e com o foco em machine learning, para aprendizagem supervisionada ou não supervisionada. (LIMA, 2020).

Fornecendo ferramentas importantes para os vários momentos do ciclo de projetos de Machine Learning, como: (ESCOVEDO, 2021).

- **Datasets:** disponibiliza alguns datasets que podem ser baixados para o projeto com poucos comandos, como o dataset Iris, um dos mais conhecidos da área de reconhecimento de padrões.
- **Pré-processamento de dados:** fornece diversas técnicas de preparação de dados, como normalização e encoding.
- **Modelos:** implementa diversos modelos de Machine Learning, tais como Regressão Linear, SVM e Random Forest, possibilitando o ajuste, avaliação e seleção do melhor modelo para o problema.

A seguir, é apresentado um resumo de suas características mais relevantes: (ESCOVEDO, 2021).

- **Consistência:** todos os objetos compartilham uma interface comum desenhada a partir de um conjunto limitado de métodos, com documentação consistente.
- **Inspeção:** todos os valores de parâmetros especificados são expostos como atributos públicos;
- **Hierarquia Limitada de Objetos:** somente algoritmos são representados por classes Python; os conjuntos de dados são representados em formatos padrão (matrizes NumPy, Pandas DataFrames, matrizes esparsas SciPy) e os nomes de parâmetros usam sequências padrão do Python;
- **Composição:** sempre que possível, as tarefas de Machine Learning são expressas como sequências de algoritmos mais fundamentais;
- **Padrões Sensíveis:** quando os modelos requerem parâmetros especificados pelo usuário, a biblioteca define um valor padrão apropriado.

O Scikit-learn apresenta uma vasta gama de itens em sua biblioteca, facilitando a aplicação de modelos, de acordo com as características da necessidade de solução relacionada.

4.5 Flask

O flask é um micro-framework do ecossistema Python, que por sua vez é uma versão minimalista de um framework tradicional, que possui uma estrutura muito mais simples e objetiva, sendo bastante utilizado para a criação de micro serviços, como por exemplo API's RESTful (API bidirecional).

Um micro-framework, é exemplificado como uma aplicação onde em sua composição total existam pequenas peças, partes "avulsas" do todo que fragmentam e facilitam a montagem da aplicação.

Dentre as principais características do Flask temos: (ANDRADE, 2020).

- **Simplicidade:** Por possuir apenas o necessário para o desenvolvimento de uma aplicação, um projeto escrito com Flask é mais simples se comparado aos frameworks maiores, já que a quantidade de arquivos é muito menor e sua arquitetura é muito mais simples.

- **Rapidez no desenvolvimento:** Com o Flask, o desenvolvedor se preocupa em apenas desenvolver o necessário para um projeto, sem a necessidade de realizar configurações que muitas vezes não são utilizadas.

- **Projetos menores:** Por possuir uma arquitetura muito simples (um único arquivo inicial) os projetos escritos em Flask tendem a ser menores e mais leves se comparados a frameworks maiores.

- **Aplicações robustas:** Apesar de ser um micro-framework, o Flask permite a criação de aplicações robustas, já que é totalmente personalizável, permitindo, caso necessário, a criação de uma arquitetura mais definida

5 Trabalhos Similares

Machine Learning and Data Mining Methods in Diabetes Research

O objetivo do presente estudo é realizar uma revisão sistemática das aplicações de aprendizado de máquina, técnicas e ferramentas de mineração de dados no campo da pesquisa em diabetes com relação a a) previsão e diagnóstico, b) complicações diabéticas, c) antecedentes genéticos e ambiente, e e) Cuidados e Gestão de Saúde com a primeira categoria aparecendo como a mais popular. Uma ampla gama de algoritmos de aprendizado de máquina foi empregada. Em geral, 85% das utilizadas foram caracterizadas por abordagens de aprendizagem supervisionada e 15% por não supervisionadas e, mais especificamente, regras de associação. (KAVAKIOTIS; TSAVE; SALIFOGLOU; MAGLAVERAS; VLAHAVAS; CHOUVARDA, 2017).

Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado

O objetivo deste trabalho, é comparar modelos de machine learning para predição de Diabetes, onde é utilizado um conjunto de dados de aproximadamente 15 mil participantes obtidos através do Estudo Longitudinal de Saúde do Adulto (ELSA – Brasil). As variáveis de referência para predição foram selecionadas para ser informações simples dos participantes, sem necessidades de exames laboratoriais. (OLIVERA, 2016).

Os testes foram realizados em quatro etapas: ajuste dos parâmetros através de validação cruzada, seleção automática de variáveis, validação cruzada para estimativa de erros e teste de generalização em um conjunto independente dos dados. Os resultados demonstram a viabilidade de utilizar informações simples para detectar casos diabetes não diagnosticado na população. Além disso, os resultados comparam algoritmos de aprendizagem de máquina e mostram a possibilidade de utilizar outros algoritmos, alternativamente à Regressão Logística, para a construção de modelos preditivos (OLIVERA, 2016).

6 Metodologia

6.1 População, coleta e amostra de dados

Os dados foram retirados do dataset público Diabetes Health Indicators Dataset disponibilizado pela plataforma Kaggle.

O dataset possui 441.455 registros, os dados implícitos subjacentes vêm do BRFSS 2015 do CDC para o artigo foram utilizadas as informações: HighBP, HighChol, BMI, Smoker, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income, contidas no diabetes_binary_5050split_health_indicators_BRFSS2015.

A coleta de dados é feita anualmente pela CDC, onde são contatados mais de 400 mil americanos sobre comportamento, hábito e estilo de vida. Esse serviço é realizado desde 1984, sendo de extrema importância para o cenário do Diabetes nos EUA.

A amostra de dados utilizada trata-se de um arquivo em formato csv, onde os dados são balanceados a 50 / 50, ou seja, 50% dos dados apontam para pessoas que de fato obtiveram a doença, e os outros 50% são de pessoas que se mantiveram saudáveis no período.

6.2 Métodos a serem utilizados

Com o dataset pré-processado foi desenvolvido um algoritmo para as técnicas de predição do resultado utilizando redes neurais: MLPClassifier (Classificador de Perceptron de várias camadas), o resultado foi apresentado através de uma API RESTful.

6.3 Ferramentas e tecnologias utilizadas

Para realização do artigo foi utilizada a plataforma Google Colab para programação na Linguagem Python. As bibliotecas utilizadas foram, Pandas para manipulação dos dados;
Scikit-Learn para criação da extração e característica e treinamento do modelo;
Flask para criação de API para interação com os dados.

7 Desenvolvimento

7.1 Pré-processamento

O dataset foi fornecido pela plataforma Kaggle, foi utilizado o arquivo balanceado 50/50 proveniente da pesquisa: Behavioral Risk Factor Surveillance System (BRFSS)

programa contínuo de vigilância de doenças crônicas por telefone projetado para coletar dados sobre os comportamentos e condições que colocam os adultos em risco de doenças crônicas, lesões e doenças infecciosas evitáveis.

7.1.1 Alteração do dataset

Em busca de alcançar a melhor precisão e reduzir a quantidade de dados, foi realizado a remoção de colunas presentes no dataset conforme aponta a imagem 1. Foram mantidas as colunas: 'HighBP', 'HighChol', 'BMI', 'Smoker', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education', 'Income' e então removidas as colunas: 'Stroke', 'HvyAlcoholConsump', 'CholCheck'.

Imagem 1 - Limpeza dos Dados

```
import pandas as pd
import time
start = time.time()
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report

dataframe = pd.read_csv('https://raw.githubusercontent.com/zlVInne/Diabetes/main/diabetes_binary_5050split_health_indicators_BRFSS2015.csv', sep=',')

"""## Limpeza dos dados """

dataframe.drop(columns=['Stroke', 'HvyAlcoholConsump', 'CholCheck'], inplace=True)
```

Fonte – Autoria Própria

7.2 Treinamento da rede neural (MLPClassifier)

MLPClassifier significa classificador Perceptron multicamada que no próprio nome se conecta a uma rede neural.

Esse modelo pode ser usado como classificador entre duas classes (por exemplo, um diagnóstico médico de positivo/negativo para uma determinada doença).

Foi realizado o treino na seguinte ordem, conforme mostra imagem 2:

- A variável X contém o dataset em que foi realizado a limpeza dos dados.
- Foi determinado que da quantidade de dados existentes em X, 30% seriam destinados ao teste final do modelo e 70% o treinamento.

● Realizou-se o treinamento da classificação entre duas classes.

Imagem 2 - Treinamento do Modelo

```
"""## Dividir e Testar Conjunto de Dados """

X = dataframe.drop(columns=['Diabetes_binary'])

y = dataframe['Diabetes_binary']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)


"""## Treinar o modelo de aprendizado de máquina """

model = MLPClassifier(hidden_layer_sizes=(110,))

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

y_pred

y_test.to_numpy()
```

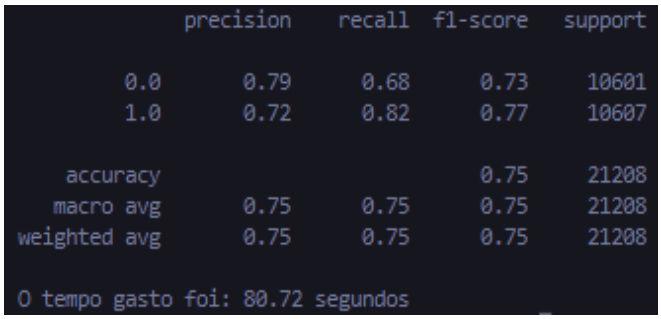
Fonte – Autoria Própria

8 Resultados e discussões

Ao elaborar esse trabalho obtivemos um tempo aproximado de 80 segundos para execução do treino e teste, conforme apurado na imagem 3 uma acurácia média de 75%, sem efetuar grandes mudanças no dataset, tampouco modificar a forma de aplicação do classificador utilizado.

Podemos combinar outros métodos para que a acurácia geral seja melhorada, trazendo então um resultado ainda mais seguro para utilização. Como podemos observar na imagem 3 tanto a acurácia de resultados positivos (1.0), quanto negativos (0.0) tem valores próximos. Para esse teste foram utilizados 21208 registros, onde podemos ver que 10601 referentes a resultados negativos (0.0) e 10607 referentes a resultados positivos (1.0), o que resulta em um dataset balanceado (50/50).

Imagem 3 - Resultados Obtidos



	precision	recall	f1-score	support
0.0	0.79	0.68	0.73	10601
1.0	0.72	0.82	0.77	10607
accuracy			0.75	21208
macro avg	0.75	0.75	0.75	21208
weighted avg	0.75	0.75	0.75	21208

O tempo gasto foi: 80.72 segundos

Fonte – Autoria Própria

Podemos trazer como discussão para esse trabalho, a hipótese da utilização dessa ferramenta no controle de Diabetes em locais onde recursos de saúde são escassos, podendo mensurar de fato a real parcela da população que necessita de tratamento especializado, melhorando assim sua qualidade de vida.

O presente dataset utilizado, não relata a realidade do Brasil, pois foi coletado a partir de dados do CDC. Tais dados se limitam a uma realidade que difere em vários aspectos da realidade encontrada no brasil. Em uma fase posterior, poderá ser coletado e utilizado dados brasileiros, a fim de trazer a realidade local, obtendo uma aplicabilidade ainda maior a realidade brasileira.

O acesso a ferramenta é livre para qualquer pessoa que possua um smartphone com conexão à internet, ou qualquer outro meio de acesso com a rede. Levando o diagnostico a lugares onde o exame com coleta de material não é simples, sendo muito impactado pela logística de materiais e profissionais de saúde.

9 Conclusão

A ferramenta elaborada possui uma grande capacidade de aplicação, devido ao fato de ser um recurso não invasivo, ou seja, que não necessita de intervenção médica para sua utilização. Serão necessários mais estudos em conjunto para que de fato se torne uma ferramenta mais completa, robusta e com dados suficientes para demonstrar nossa realidade.

Recomenda-se que os resultados sejam apurados por um corpo clínico a fim de eliminar ressalvas, e mensurar a real maneira de aplicar ao Brasil.

Podemos combinar tecnologias de detecção de doenças para acentuar a melhoria da acurácia geral, trazendo maior confiabilidade e qualidade no resultado.

A agilidade demonstrada pela ferramenta também é outro ponto que devemos ressaltar, devido ao fato que exames de diagnostico a partir de coleta de sangue, demoram a ter se resultado emitido.

Referências

ANDRADE, Ana Paula de. O que é Flask? veja neste artigo o que é o flask, principal micro-framework do ecossistema python.. Veja neste artigo o que é o Flask, principal micro-framework do ecossistema Python. 2020. Disponível em: https://www.treinaweb.com.br/blog/o-que-e-flask. Acesso em: 01 jun. 2022.

BRASIL. Escola Superior da Tecnologia da Informação. Instituto Infnet (org.). Vamos falar de Python. 2020. Disponível em: https://www.infnet.edu.br/esti/vamos-falar-de-python/. Acesso em: 31 maio 2022.

BRASIL. MINISTÉRIO DA SAUDE. 26/6 – Dia Nacional do Diabetes. 2020. Disponível em: https://bvsms.saude.gov.br/26-6-dia-nacional-do-diabetes-4/. Acesso em: 19 mar. 2022.

BRASILIA. Ministério da Saúde. Coordenação Nacional do Plano de Reorganização da Atenção À Hipertensão Arterial - Ha e Ao Diabetes Mellitus - Dm. Plano de Reorganização da Atenção à Hipertensão arterial e ao Diabetes mellitus: manual de hipertensão arterial e diabetes mellitus. Manual de Hipertensão arterial e Diabetes mellitus. 2002. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/miolo2002.pdf. Acesso em: 19 mar. 2022.

CORRêA, Eduardo. Pandas Python. 2020. Disponível em: https://www.casadocodigo.com.br/pages/sumario-pandas-python. Acesso em: 31 maio 2022.

ESCOVEDO, Tatiana. Implementando um Modelo de Classificação no Scikit-Learn: o que é o scikit-learn? O que é o Scikit-Learn? 2021. Disponível em: https://tatianaesc.medium.com/implementando-um-modelo-de-classifica%C3%A7%C3%A3o-no-scikit-learn-6206d684b377. Acesso em: 31 maio 2022.

KAVAKIOTIS, Ioannis; TSAVE, Olga; SALIFOGLOU, Athanasios; MAGLAVERAS, Nicos; VLAHAVAS, Ioannis; CHOUVARDA, Ioanna. Machine Learning and Data Mining Methods in Diabetes Research. 2017. Disponível em: https://pdf.sciencedirectassets.com/311228/1-s2.0-S2001037016X00025/1-s2.0-S2001037016300733/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEHcaCXVzLWVhc3QtMSJHMEUCIFb3BR2JkA5IwjXD2dV%2FyNZ59J7JtOWh3lvd1uWej%2FxVAiEAyndVfnIMO8a7glgPlKE1T5Db7Ho2a7mOcjRX2L

RDA30qgwQI8P%2F%2F%2F%2F%2F%2F%2F%2F%2FARAEGgwwNTkwMDM1NDY4NjUiDNdV6Q2KH3ze2xt6PCrXA3ry0UhKMyiBHBgtYMLYFeNkmgGd8e%2FMB7YDaNfj%. Acesso em: 31 maio 2022.

LIMA, Guilherme. Python: A origem do nome. 2021. Disponível em: https://www.alura.com.br/artigos/python-origem-do-nome?gclid=Cj0KCQjw-daUBhCIARIsALbkjSZgFp75lfmFm3J5xrKLDaLJBxGVceB4FmVcTq2DG_G9mjL4tv0etCEaAhLREALw_wcB. Acesso em: 31 maio 2022.

LIMA, Rodrigo. Competição DSA de Machine Learning: competição dsa de machine learning - edição dezembro/2019. Competição DSA de Machine Learning - Edição Dezembro/2019. 2020. Disponível em: https://www.kaggle.com/c/competicao-dsa-machine-learning-dec-2019/discussion/121477. Acesso em: 20 mar. 2022.

OLIVERA, André Rodrigues. Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado. 2016. 84 f. Dissertação (Mestrado) - Curso de Mestre em Ciencia da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

RIBEIRO, Thiago. Kaggle – O que é? Como Funciona? 2018. Disponível em: https://tirandolicoesdetudo.com.br/kaggle-o-que-e-como-funciona/. Acesso em: 31 maio 2022.

SAHU, Mariane. Machine Learning na Saúde: prevenção, detecção e tratamento de doenças. 2020. Disponível em: https://www.unisoma.com.br/machine-learning-na-saude-prevencao-tratamento/. Acesso em: 19 mar. 2022.

TENORIO, Goretti; PINHEIRO, Cholé. O que é diabetes tipo 2: causas, sintomas, tratamentos e prevenção Leia mais em: https://saude.abril.com.br/medicina/o-que-e-diabetes-tipo-2-causas-sintomas-tratamentos-e-prevencao/: a versão mais comum do diabetes está ligada aos hábitos de vida (obesidade, sedentarismo, alimentação inadequada). conheça a doença e suas consequências. A versão mais comum do diabetes está ligada aos hábitos de vida (obesidade, sedentarismo, alimentação inadequada). Conheça a doença e suas consequências. 2019. Disponível em: https://saude.abril.com.br/medicina/o-que-e-diabetes-tipo-2-causas-sintomas-tratamentos-e-prevencao/. Acesso em: 05 jul. 2022.

WIKIPÉDIA. Python. 2022. Disponível em: https://pt.wikipedia.org/wiki/Python. Acesso em: 20 mar. 2022.