Introdução

A diabete está entre as doenças crônicas mais prevalentes nos Estados Unidos da América (EUA), impactando milhões de Americanos a cada ano exercendo um encargo financeiro significativo na economia. É uma doença crônica grave em que os indivíduos perdem a capacidade de regular níveis de glicose (acuçar) no sangue, podendo levar à redução de qualidade de vida e da expectativa de vida (Teboul, A., 2021).

A diabete geralmente é caracterizada pelo facto do corpo não produzir insulina suficiente ou ser incapaz de usar insulina de forma tão eficaz quanto necessário, o que pode provocar complicações como: doenças cardíacas, perda de visão, amputação de membros inferiores e doença renal.

O diagnóstico precoce pode levar a mudanças no estilo de vida e tratamento mais eficaz, tornando os modelos preditivos de risco de diabetes ferramentas importantes para autoridade públicas e de sáude pública.

Objetivos do projeto

Neste projeto cobrimos todas as etapas de um projeto real de Ciência de Dados e respondemos a algumas questões importantes sobre a área de saúde (diabetes), utilizando dados disponibilizados no kaggle sobre uma pesquisa feita nos EUA, com intuito de permitir que tenhamos conhecimento e/ou descubramos insigths que não estão evidenciados de forma explicita sobre diabetes que é uma doença que temos casos no nosso dia a dia:

- Qual é a corelação entre as avriáveis preditoras e a variável alvo?
- Qual é a distribuição dos entrevistados por cada classe?
- Qual é a distribuição das variáveis índice de massa corporal, idade, nível de pessão arterial, nível de colestrol é igual em cada uma das classes?
- Quais os factores que mais influenciam na obtenção de diabetes?

O objetivo é dar a conhecer alguns *insigths* extraídos a partir dos dados e criar uma solução (modelo preditivo e dashboard) para a área de saúde que permite classificar um paciente de forma precoce nas seguintes classes: **diabético, pré-diabético e não diabéticos**.

Por fim, realçar que o grande impacto que o projeto terá na área de saúde é o que a seguir se descreve:

- Esta ferramenta e/ou solução tecnológica permitirá que os orgãos responsáveis pela saúde pública investiguem com mais detalhes o que contribui e/ou quais os factores que influenciam nesta doença, de modo a definir-se estratégias a curto prazo para identificação precoce de diabetes (no estágio de pré-diabetes) e permitindo um tratamento mais eficaz.

Pré-diabetes é uma condição de saúde grave em que os **níveis de acuçar no sangue são mais altos do que o normal**, mais ainda não altos o suficiente para serem diagnosticados como diabetes do tipo 2 (Wikipédia).

Solução Proposta

Tecnologias Utilizadas

Para resolver este problema foi construida uma solução completa para armazenamento e gestão usando **Google Cloud Plataform** (GCP), além de explorar uma suite de tecnologias e/ou bibliotecas para análise, visualização de dados e *machine learning* tais como: **pandas, matplotlib, seaborn, scikit-learn, streamlit, pycaret e pyspark.**

Pandas – biblioteca usada para manipulação de dados

Matplotlib – biblioteca usada para visualização de dados.

Seaborn – biblioteca usada para visualização de dados baseada no matplotlib, permitindo construir graficos mais profissonais.

Scikit-learn – biblioteca usada para implementar os algoritmos de *machine learning* (utilizou-se o pickle para serializar o modelo em disco)

Pycaret – biblioteca open-source usada para fazer Auto-ML em um projeto de ciência` de dados.

Streamlit – biblioteca utilizada para desenvolver a aplicação e/ou fórmulário para testar o modelo em ambiente de produção.

Ferramentas de auxiliares:

Pyspark – para processamento de grandes volumes de dados em ambientes distribuído.

Python – liguagem de programação utilizada para desenvolver o projeto de ciência de dados.

Google colab – editor de código online que geralmente é organizado por células que permite executar todas as etapas de um projeto de ciência de dados.

Github – ferramenta que permite versionar, partilhar o código desenvolvido e também atribuir acesso a outros profissinais para colaborarem nos artefatos do projeto.

Apache Airflow - é uma plataforma de gerenciamento de fluxo de trabalho de código aberto para pipelines de engenharia de dados.

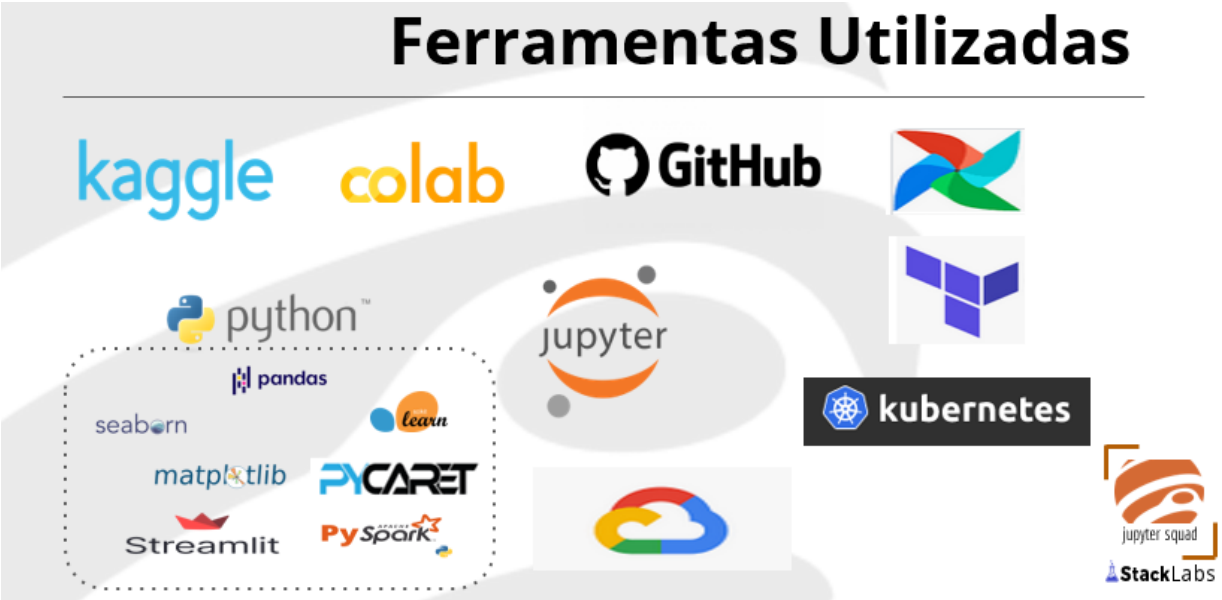
Terraform – é uma ferramenta do tipo infraestrutura como código (IaC) que permite o gerenciamento e provisionamento da infraestrutura por meio de códigos, em vez de

processos manuais. Esta ferramenta foi utilizada para criar os buckets de forma automatizada.

Kubernetes - é uma ferramenta para orquestrar os serviços disponibilizadas no docker, ou seja, é um plataforma de código aberto, portável e extensiva para o gerenciamento de cargas de trabalho e serviços distribuídos em contêineres, que facilita tanto a configuração declarativa quanto a automação

Google Cloud Plataform (GCP) - é uma suíte de computação em nuvem oferecida pelo Google, funcionando na mesma infraestrutura que a empresa usa para seus produtos dirigidos aos usuários

Overview Geral de Tecnologias Utilizadas no Projeto



Arquitecturas

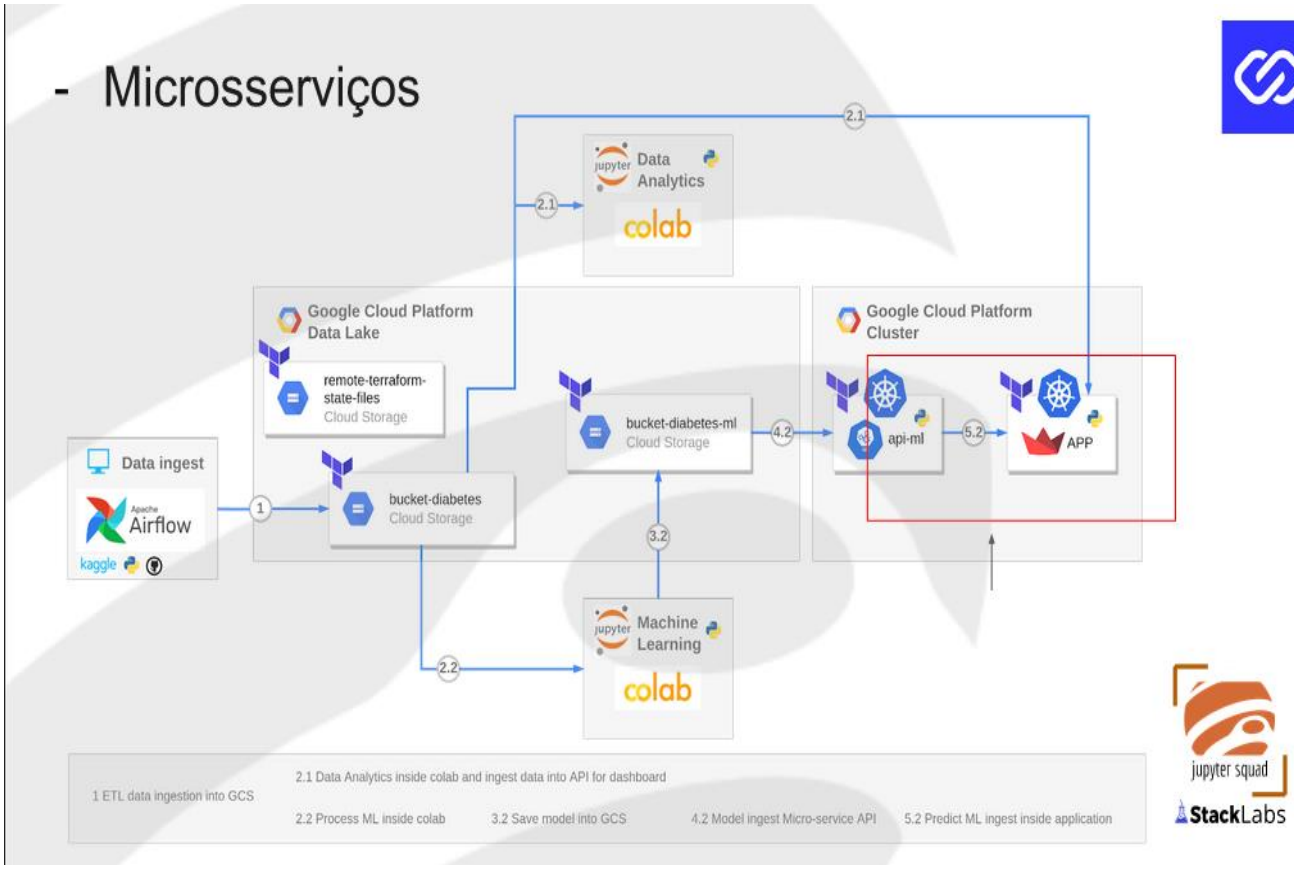
Em seguida é ilustrada o overview da solução desde a **coleta** até ao **deploy** da solução desenvolvida**.**

Projetada pela squad **Jupyter** cujos os integrantes são:

- Pedro Lucas – Data Analyst and Project Leader

- Celso Adamo - Data Scientist

- Adilson Silva - Data Engineer



Os principais desafios enfrentados foram:

- **Integrar** o notebook do Google Colab com o GCS usando emails pessoais, permitindo desta forma a leitura dos datasets armazenados no GCP.
- **Carregar** o modelo na app do streamlit devido a incompatibilidade de versões.
- **Criar** a api usando o framework web FastAPI.
- **Conectar** o GCS com api-ml e a aplicação.

Resultados

Insights e Conhecimento Gerado

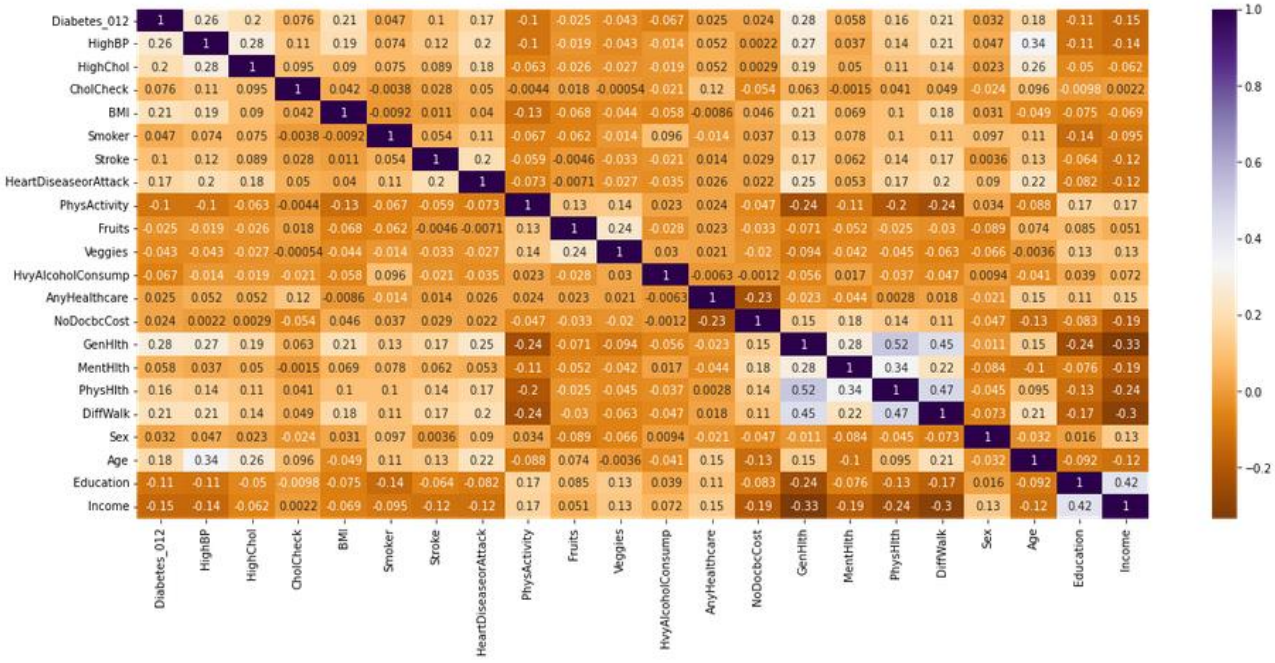
Na etapa de Análise Exploratória dos Dados foram descobertos vários insights importantes abaixo descritas.

Pela análise estatistica básica feita sobre os dados apurou-se o seguinte:

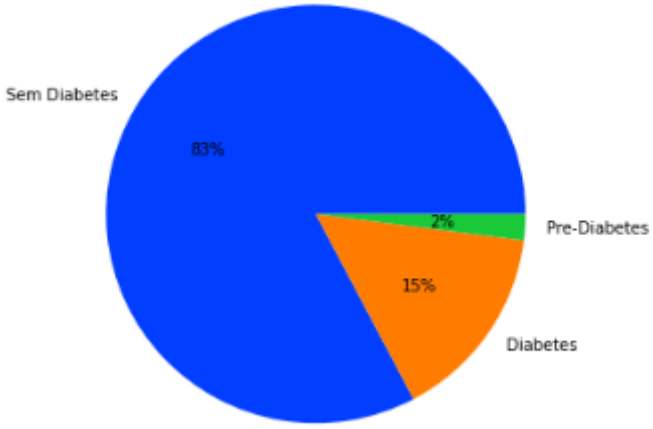
- Em médias os entrevistados têm um **índice de massa corporal (BMI)** de **28.68**.
- Quase **metade** das pessoas entrevistadas **fumam e/ou comem fru**ta.
- **73.34 %** das pessoas entrevistadas praticam **actividades físicas**.
- **79.48 %** das pessoas entrevistadas comem **vegetais**.
- Quase **ninguém** consume **álcool em altas proporções** (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week).
- **94%** das pessoas entrevistadas usaram algum **plano e/ou seguro de saúde.**

Foram feitas algumas questões sobre os dados e constatamos o seguinte:

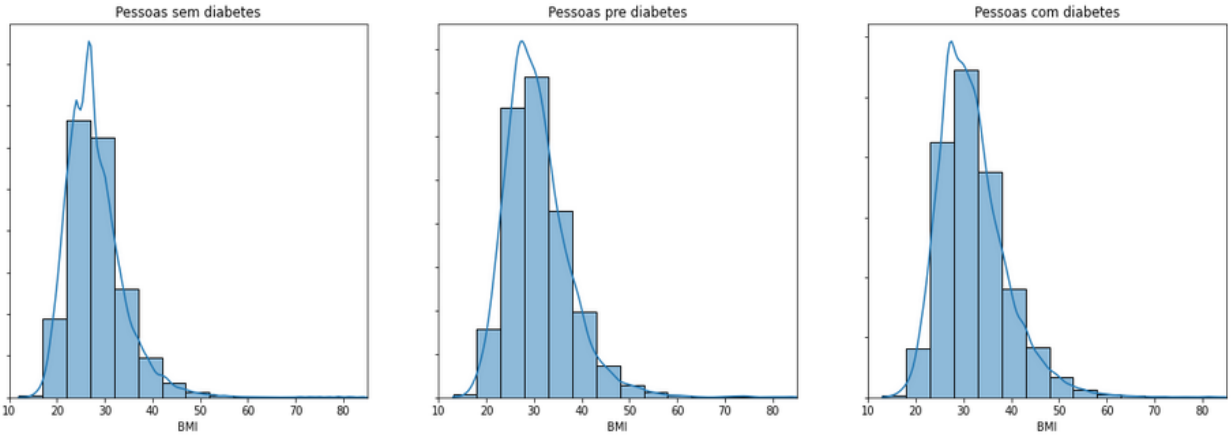
A maioria das variáveis possuem uma **correlação fraca** entre elas exceptuando a variáveis **PhysHlthe GenHlth** que possuem uma **correlação média**.



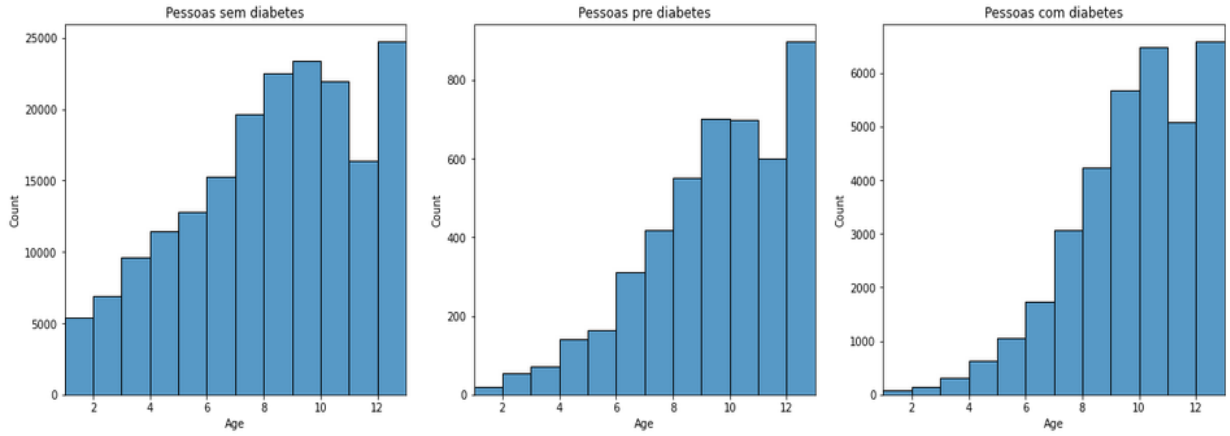
Analisando a distribuição dos entrevistados por classes constatou-se que **83%** dos entrevistados **não tem diabetes**.



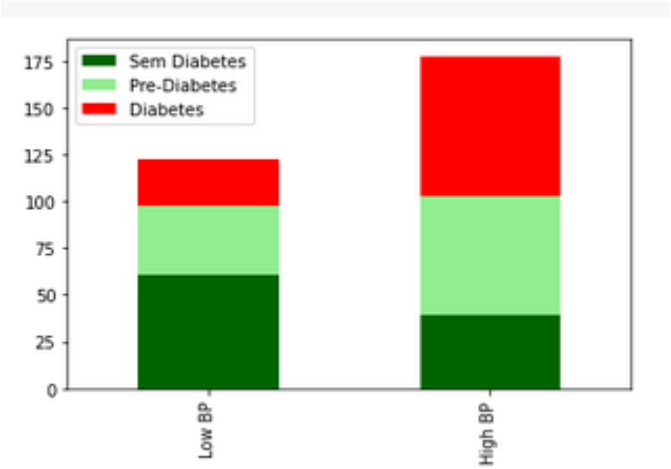
O índice de massa coporal tem o mesmo dominio de valores para as 3 classes envolvidas.

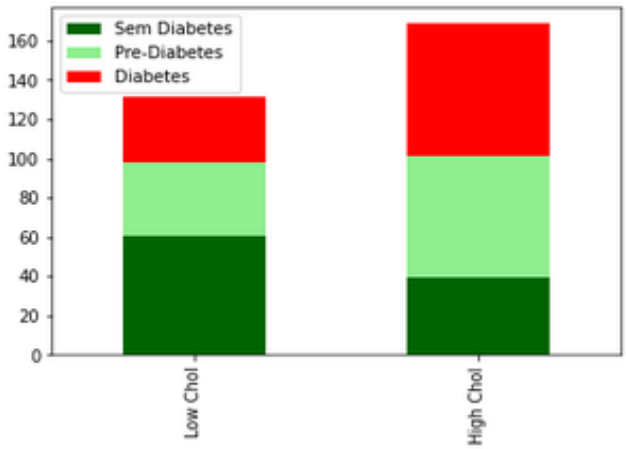


A maior parte das pessoas com **pré-diabetes e diabetes** estão em **idade** pertencentes a **categoria 6 em diante**. Não obstante, não chegamos a nenhuma conclusão de grau de obtenção de diabetes em função da idade.



A **maior parte das pessoas** com **pressão arterial alta** e/ou **colesterol alto** pertencem as categorias **pré-diabetes e diabetes**.





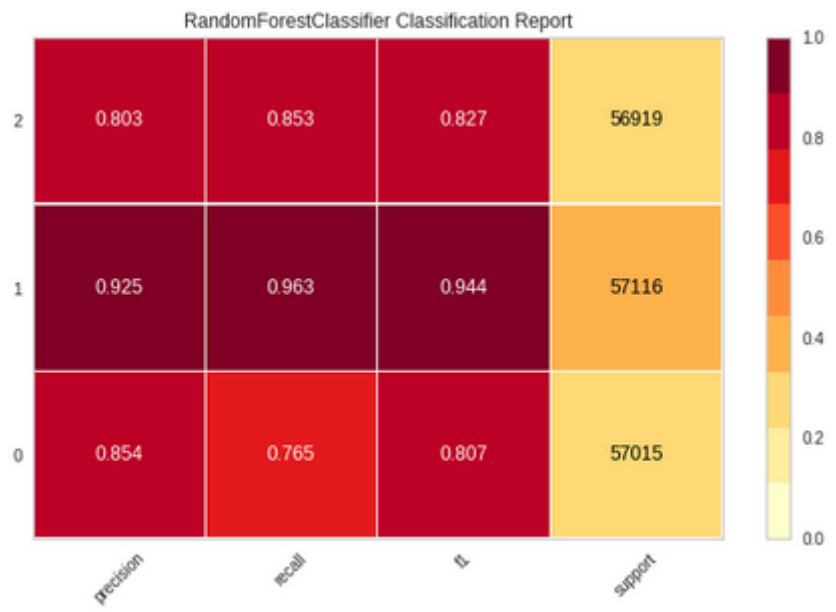
Grau de importâncias das variáveis preditoras no modelo em percentagem

BMI	18.426225
Age	12.333248
Income	10.112030
PhysHlth	8.529593
Education	7.265321
GenHlth	6.537013
MentHlth	6.362040
HighBP	3.627002
Fruits	3.478912
Smoker	3.473712
Sex	2.905548
Veggies	2.750739
PhysActivity	2.744945
HighChol	2.641753
DiffWalk	2.206154
HeartDiseaseorAttack	1.790474
NoDocbcCost	1.541153
Stroke	1.192858
AnyHealthcare	0.894155
HvyAlcoholConsump	0.799575

Métricas de Performance

Para predizer se o paciente pertence a cada uma das classe foi implementado um modelo utilizando o **Random Forest** que atingiu uma performance **F1-Score** de aproximadamente **85%** (84.70% - superando um pouco um modelo de base que é de **75%**).

Como tratava-se de um problema de classes desbalanceadas e houve necessidade utilizar a técnica de *over sampling* denominada **SMOTE** e de analisar as métricas **precision** e **recall** como forma de analisar a performance de previsão de cada uma das classes.



Conclusão

Através deste projeto foi possível praticar e implementar conceitos importantes de Ciência e Engenharia de Dados e propor uma solução para a área de saúde que permite **descobrir alguns *insigths*** sobre diabetes e **classificar se um paciente** pertence a cada uma das classes identificadas na secção do objectivo.

A resolução dos problemas acima mencionados, permitirá obter insights para desenhar e/ou alterar a estratégia aplicada á área de saúde no que concerne a diabetes, visto que, esta doença tem um impacto significativo na economia e pode ter complicaçoes quando nao detectado em estágios mais precoces.

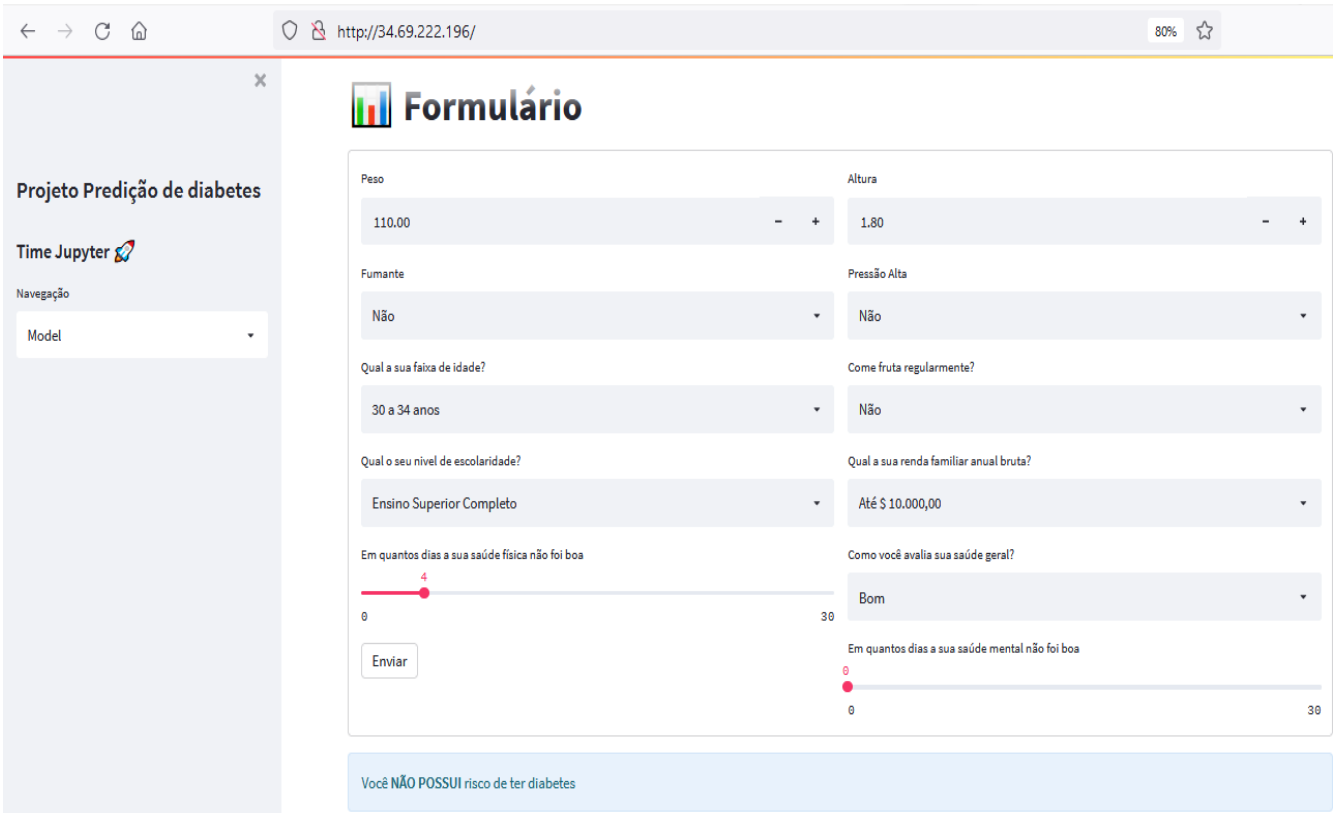
Com a implementação desta solução teremos como beneficio, um recurso organizacional que servirá de apoio aos médicos na leitura de análises médicas e contribuirá para identificar as principais variáveis que influenciam em cada uma das classes.

Por fim, como um processo de melhoria continua pode-se reduzir o erro de previsão do modelo utilizando outras técnicas como *feature engeneering*, redução de dimensionaldade, entre outras e criar mais interatividade no dashboard do streamlit.

Anexos

Link do Streamlit e dashboard: http://34.69.222.196/

App



Dashboard