Projeto Predição de Diabetes

🚀 Time Jupyter - StackLabs

- Adilson Gomes da Silva Junior - Engenheiro de dados
- Celso Meirelles Rodolfo Adamo - Cientista de dados
- Pedro Lucas Grajaú Farias - Analista de dados

jupyter squad

StackLabs

Base de dados - Kaggle

- Diabetes Health Indicators Dataset - https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
- A pesquisa é feita anualmente com mais de 400 mil entrevistados nos EUA e com cerca de 330 características relacionadas à saúde
- Dataset do tratado no Kaggle possui aproximadamente 250 mil entrevistados (linhas) e 21 características (colunas) com foco em diabetes

jupyter squad

StackLabs

Objetivos

- Qual é a correlação entre as variáveis preditoras e a variável alvo?
- Quais os principais fatores que influenciam a ocorrência de casos de diabetes?
- Existe algum fator que predomina sobre os outros?

jupyter squad

StackLabs

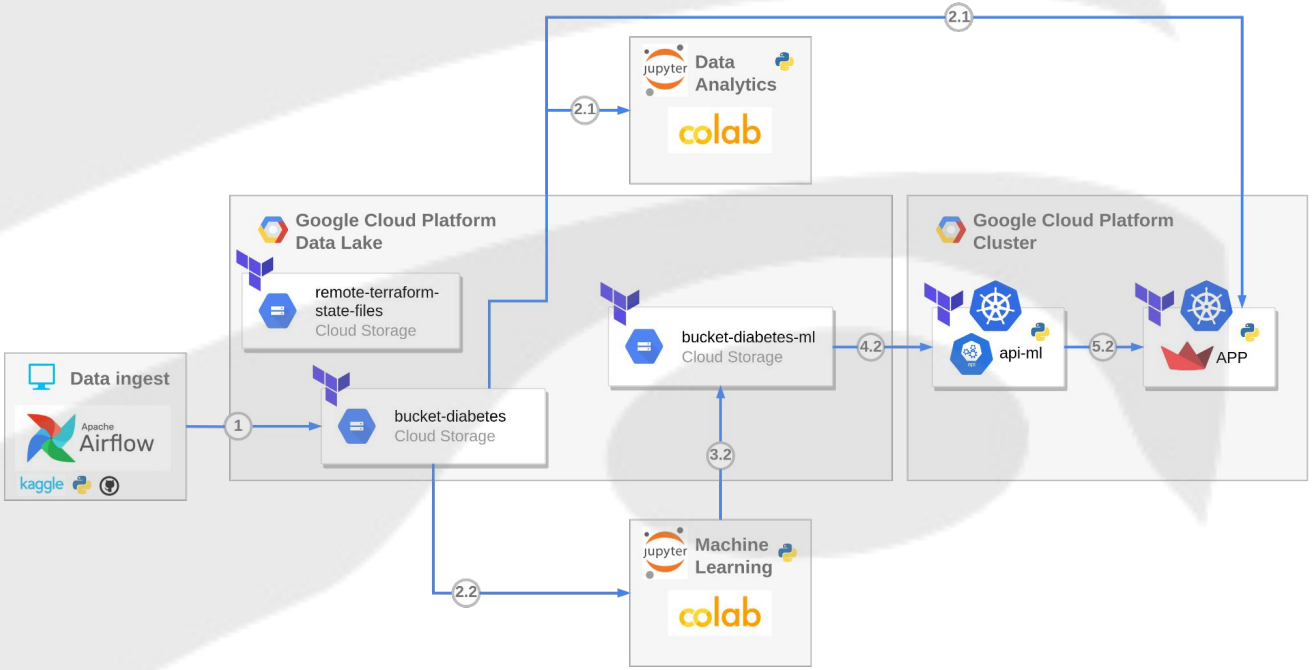
Definições do projeto

- Plataforma - cloud
- Armazenamento de código
- Ferramenta IaC
- Linguagem de programação
- Plataforma para análise e machine learning
- Ferramentas de orquestração
- Ferramenta de visualização

jupyter squad

StackLabs

Pipeline de dados



jupyter squad

StackLabs

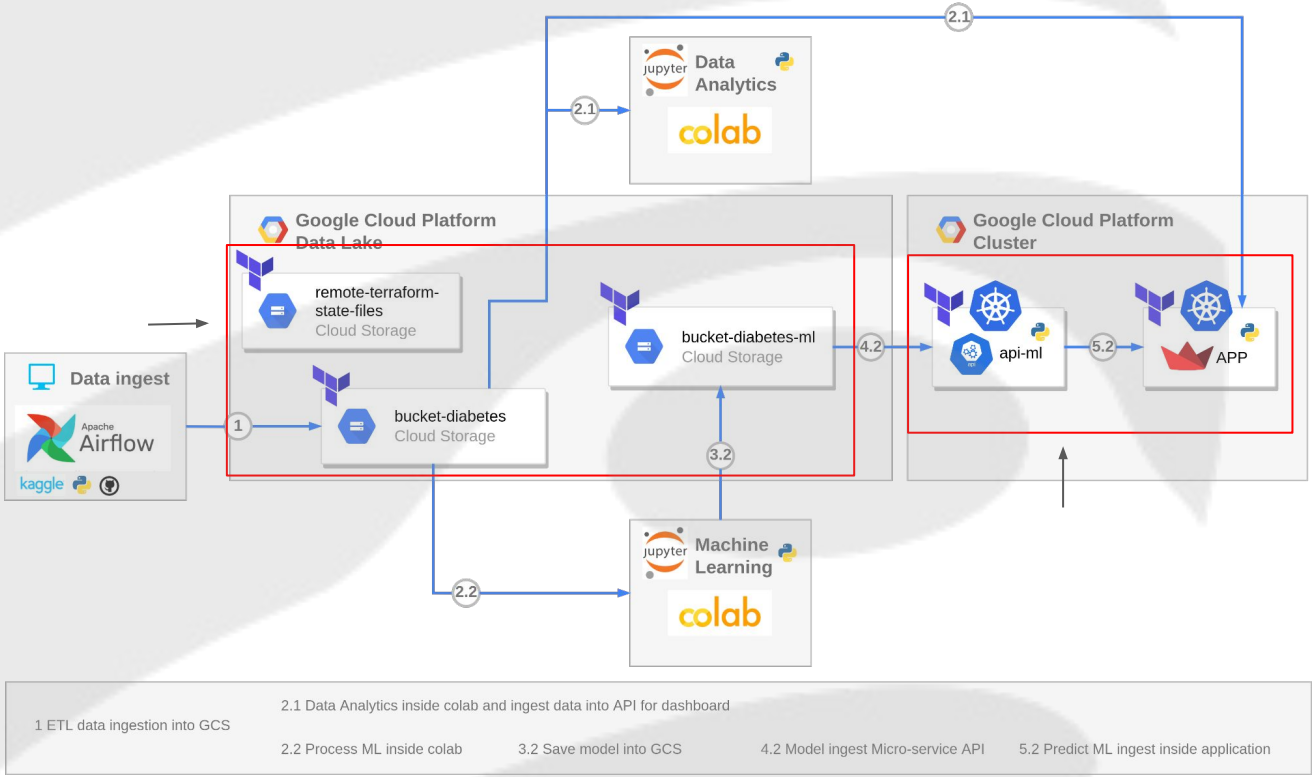
ETL dos dados

1 Get .csv data on kaggle 2 - Clean data - Spark submit Operator (local) 3 Save .parquet data into GCS

jupyter squad

StackLabs

Micro-serviço Data lake e Cluster



jupyter squad

StackLabs

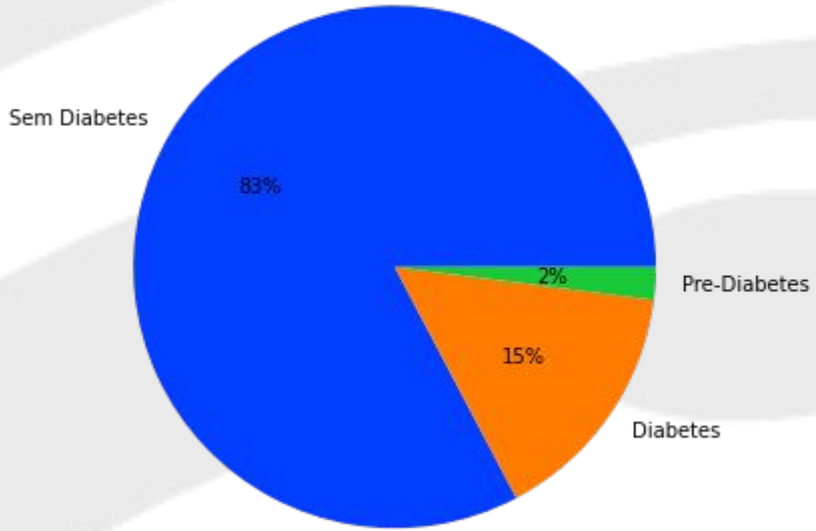
Principais desafios

- Conexão Google colab com o GCS
- Conexão GCS com a api-ml e a aplicação - Montar um volume com o driver GCSFuse
- Conectar o modelo com a aplicação - Incompatibilidade de versões
- Criação da api - Utilização do framework web FastAPI

jupyter squad

StackLabs

Insights e Conhecimento Gerado



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

jupyter squad

StackLabs

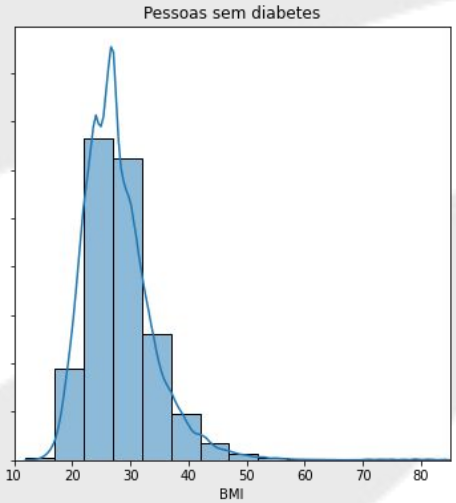
Insights e Conhecimento Gerado

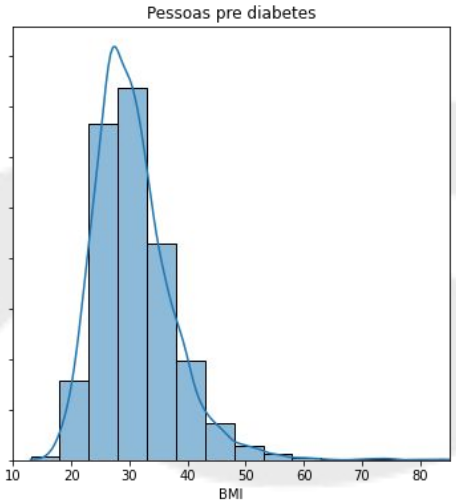
- Em médias os entrevistados têm um **índice de massa corporal (BMI)** de **28.68**.
- **73.34 %** das pessoas entrevistadas praticam **atividades físicas**.
- **79.48 %** das pessoas entrevistadas comem **vegetais**.
- Quase **ninguém** consome **álcool em altas proporções** (adult men having more than 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.**

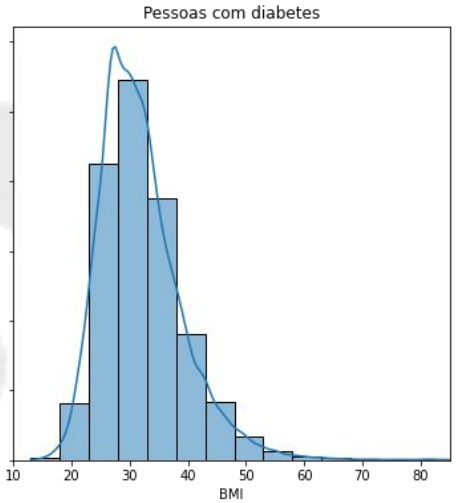
jupyter squad

StackLabs

Insights e Conhecimento Gerado





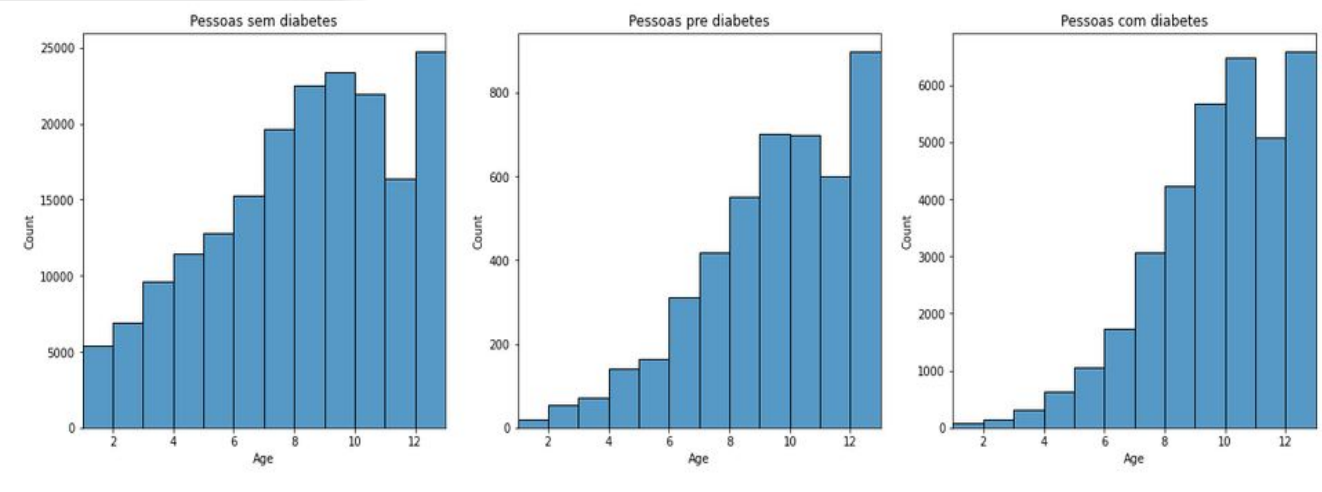


Podemos ver uma relação de casos de diabetes com o aumento de IMC, principalmente entre os bins de 30 e 40.

jupyter squad

StackLabs

Insights e Conhecimento Gerado

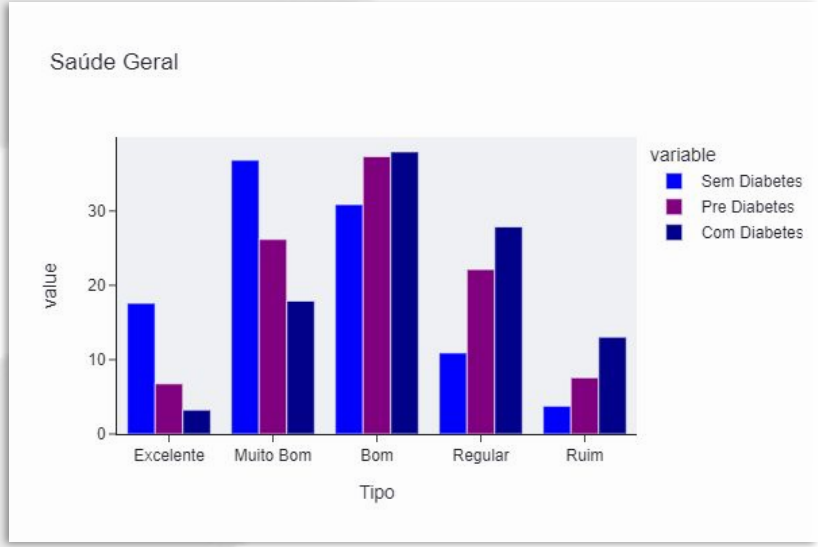


Podemos ver uma relação também relacionada à idade, visto que há uma concentração bem maior de pessoas com idades mais avançadas com Diabetes.

jupyter squad

StackLabs

Insights e Conhecimento Gerado



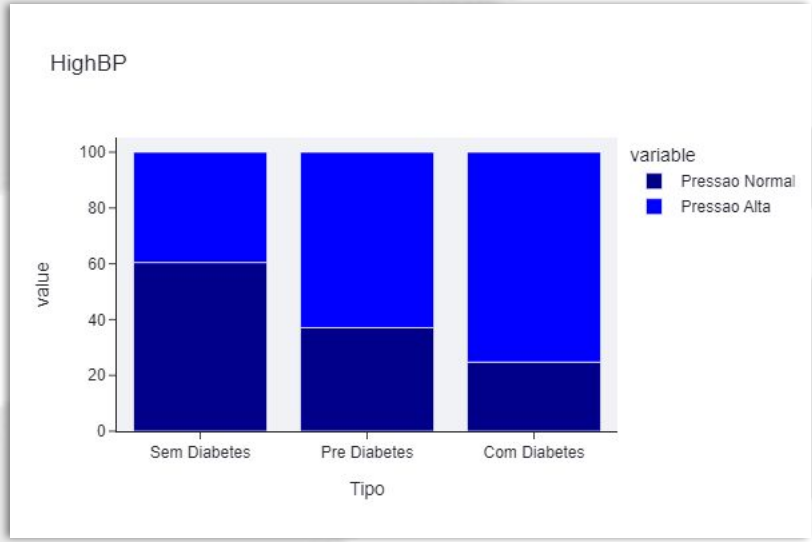
Como é esperado, a questão diabetes afeta diretamente com a saúde geral.

Podemos ver uma taxa maior de saúde excelente ou muito bom para pessoas com diabetes e também uma taxa maior de regular e ruim para pessoas com diabetes.

jupyter squad

StackLabs

Insights e Conhecimento Gerado



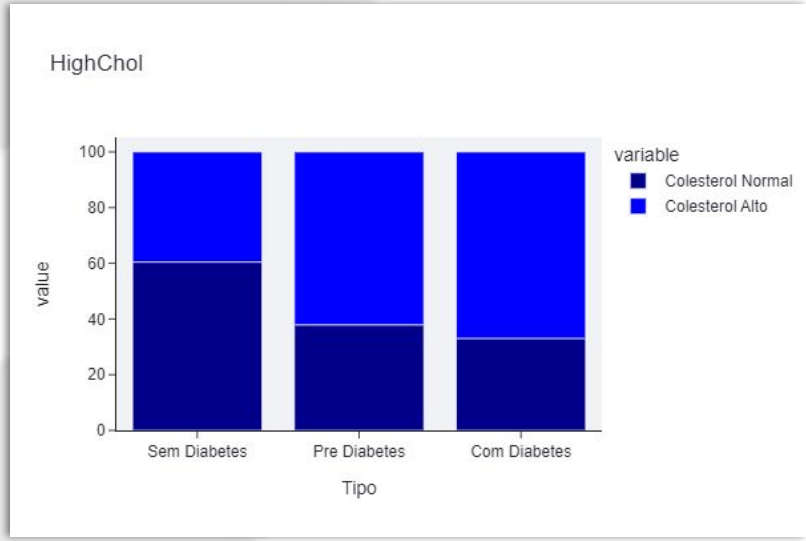
Agora falando de variáveis categóricas, podemos ver uma variável bastante relevante que é a de Pressão Alta.

Podemos ver que há uma relação bastante forte com pessoas com diabetes, **sendo 75% de pessoas com diabetes tem também pressão alta.**

jupyter squad

StackLabs

Insights e Conhecimento Gerado



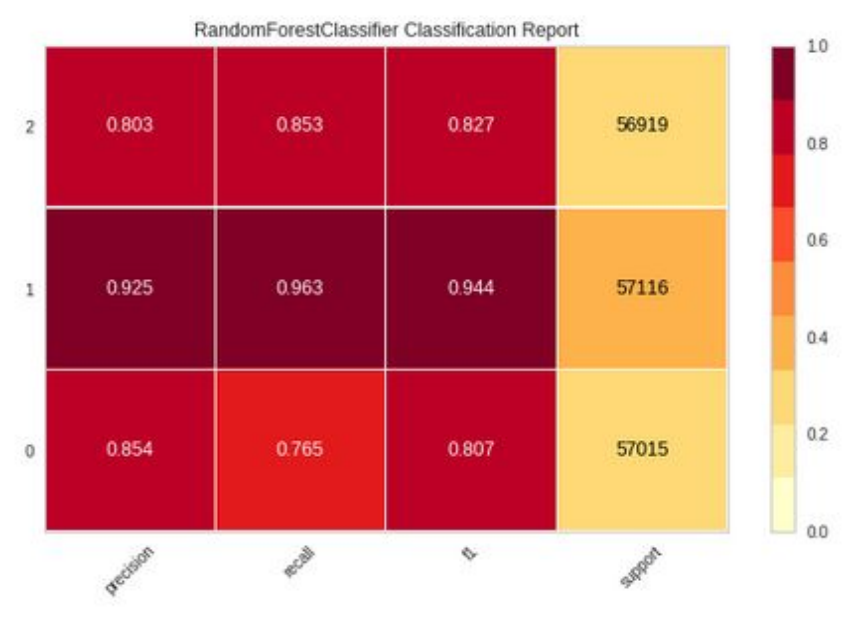
Assim como o gráfico anterior, há uma relação bastante relevante também relacionada a diabetes.

Temos **66% de pessoas com colesterol alto e com diabetes.**

jupyter squad

StackLabs

Métricas de Performance

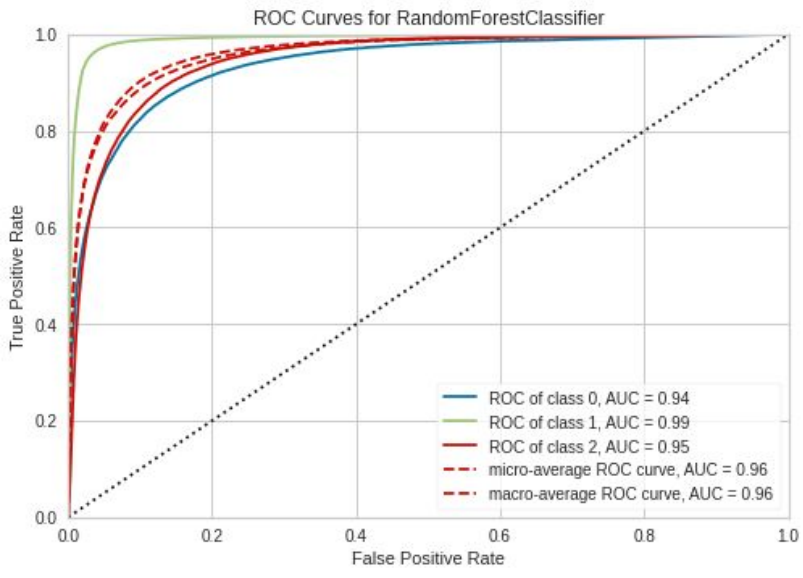


Para predizer se o paciente pertence a cada uma das classe foi implementado um modelo utilizando o Random Forest que atingiu uma performance F1-Score de **aproximadamente 85% (84.70% - superando um pouco um modelo de base que é de 75%)**

jupyter squad

StackLabs

Métricas de Performance

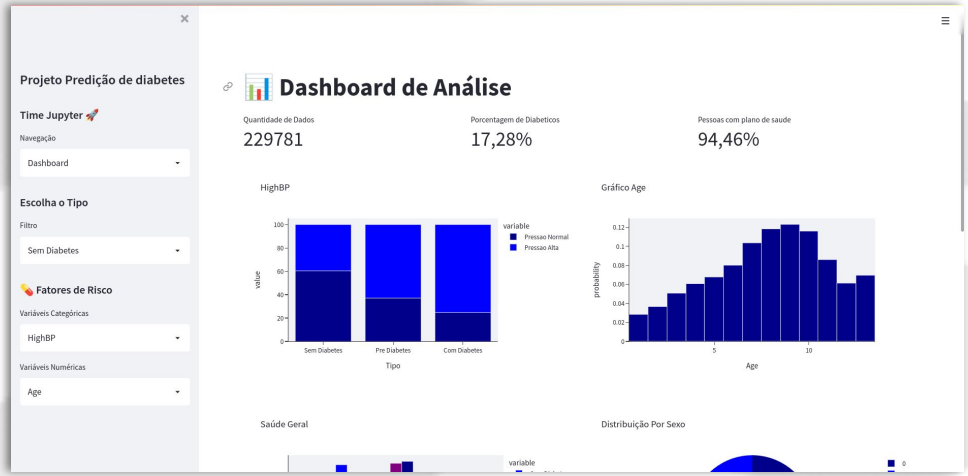


Tratava-se de um problema de classes desbalanceadas e houve necessidade utilizar a técnica de over sampling denominada SMOTE e de analisar o precision e recall como forma de analisar a performance de previsão de cada uma das classes.

jupyter squad

StackLabs

Entregáveis



Como entregável decidimos por entregar duas partes, uma relacionada a o **dashboard** com as informações encontradas no dataset e outro relacionado a um **formulário** em que o usuário irá colocar os seus dados e haverá uma predição e um score dessa predição.

jupyter squad

StackLabs

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 insights sobre diabetes e classificar se um paciente pertence a cada uma das classes.

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

Com a implementação desta solução teremos como benefício, 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 contínua podemos reduzir o erro de previsão do modelo utilizando outras técnicas como feature engineering, redução de dimensionalidade, entre outras.

jupyter squad

StackLabs

AGRADECIMENTOS

- Adilson Gomes da Silva Junior - Data Engineer
  - https://www.linkedin.com/in/adilson-silva-junior
  - https://github.com/AdilsonSilvaJr/jupyterstack
- Celso Meirelles Rodolfo Adamo - Data Scientist
  - https://www.linkedin.com/in/celso-adamo-48773356
  - https://github.com/celsoadamo/Projeto_Diabetes
- Pedro Lucas Grajaú Farias - Data Analyst
  - https://www.linkedin.com/in/p217
  - https://github.com/Pedro-Grajau/jupyterstack

jupyter squad

StackLabs