
Faculdade
XPe
RELATÓRIO
PROJETO
APLICADO
PÓS-GRADUAÇÃO

XP Educação

Relatório do Projeto Aplicado

PERSONALIZAÇÃO DE PRÊMIOS DE SEGURO COM BASE NA QUALIDADE DO SONO

Marcelo Bin Resende da Silva

Orientador(a):

Davidson Oliveira

06/11/2023

MARCELO BIN RESENDE DA SILVA

XP EDUCAÇÃO

RELATÓRIO DO PROJETO APLICADO

PERSONALIZAÇÃO DE PRÊMIOS DE SEGURO COM BASE NA QUALIDADE DO SONO

Relatório de Projeto Aplicado desenvolvido para fins de conclusão do curso **Pós-graduação em Data Science & Machine Learning.**

Orientador (a): Davidson Oliveira

Uberlândia - MG

06/11/2023

Sumário

1. CANVAS do Projeto Aplicado

Figura conceitual, que representa todas as etapas do Projeto Aplicado.

1.1 Desafio

1.1.1 Análise de Contexto

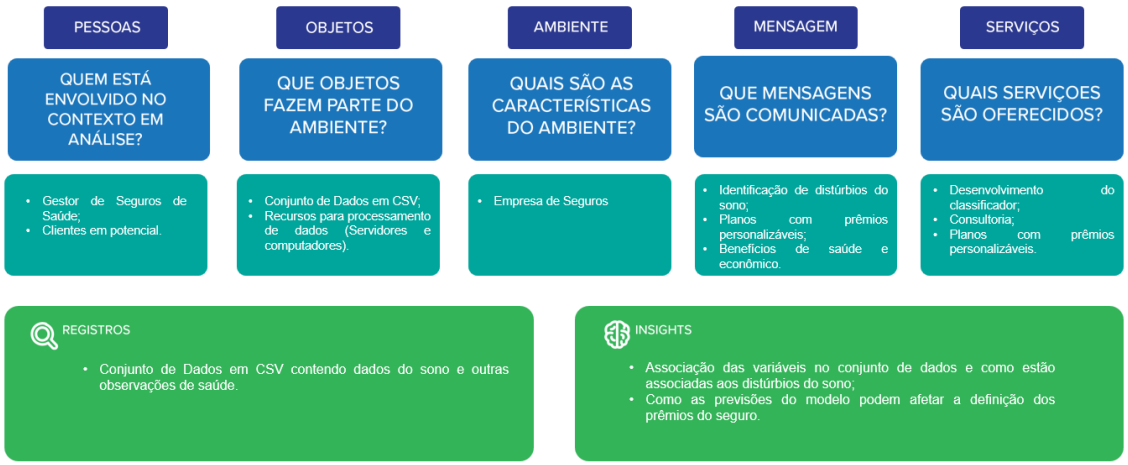
O CEO da SAFE Seguros, Sr. Gilberto Safe, lidera uma seguradora especializada em saúde há 15 anos. A missão da empresa é "Proteger a dignidade de nossos segurados em todos os momentos de sua vida".

Recentemente, o Sr. Gilberto participou de um Congresso de Saúde que destacou o impacto da qualidade do sono na saúde. Estudos indicam que distúrbios do sono podem contribuir para problemas de saúde mais sérios.

Ele adquiriu um estudo realizado por uma clínica de sono parceira de sua seguradora, que monitorou a qualidade do sono de 400 voluntários. Agora, ele busca nossa consultoria para aproveitar esses dados e criar um modelo de classificação que permita aprimorar os seguros de saúde da empresa, ajustando os prêmios com base na classificação do tipo de sono.

O seguro saúde da SAFE Seguros oferece assistência médica, consultas, exames, cirurgias, internações e tratamentos. Os segurados têm a liberdade de escolher profissionais e hospitais, diferentemente dos planos de saúde que impõem restrições. Esta abordagem visa não apenas aprimorar os produtos da empresa, mas também a experiência do cliente em potencial.

Para fornecer uma visão geral das informações colhidas, utilizamos o quadro abaixo com a estrutura do POEMS:



Para nos ajudar a compreender as bases e áreas que requerem mais investigação, utilizamos a Matriz CSD abaixo. Que indica elementos de Certeza, Suposições e Dúvidas no projeto:

1.1.2 Personas

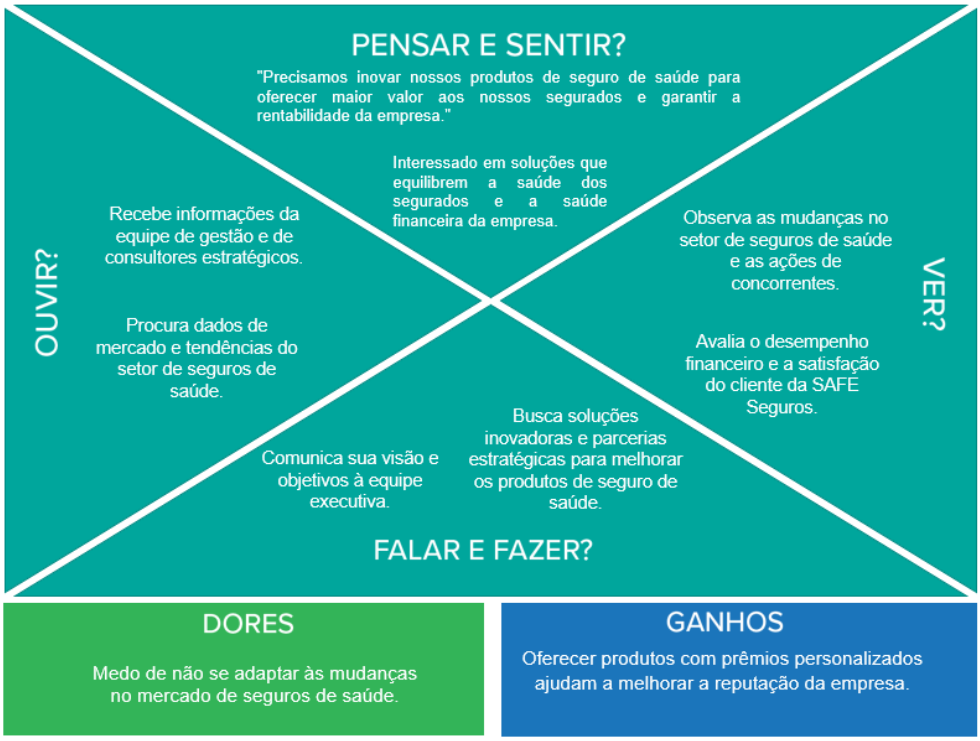
No desenvolvimento de nosso projeto, é essencial compreender as pessoas que serão impactadas por nossas decisões e estratégias. Para isso, usaremos duas pessoas que representam grupos-chave de interessados no contexto de nossa consultoria para a SAFE Seguros.

Gestor de Seguros: Sr. Gilberto Safe

Nome: Sr. Gilberto Safe
Idade: 50 anos
Profissão: CEO da SAFE Seguros
Localização: Sede da empresa, São Paulo, Brasil
Características Comportamentais:

- Focado na rentabilidade e no crescimento da empresa.
- Interessado em inovações que melhorem a oferta de seguros de saúde.
- Preocupado com a satisfação e a proteção dos segurados.
- Aberto a parcerias estratégicas para alcançar os objetivos da empresa.

Mapa de Empatia de Gilberto Safe:



Cliente em Potencial: Ana Silva

Nome: Ana Silva

Idade: 32 anos

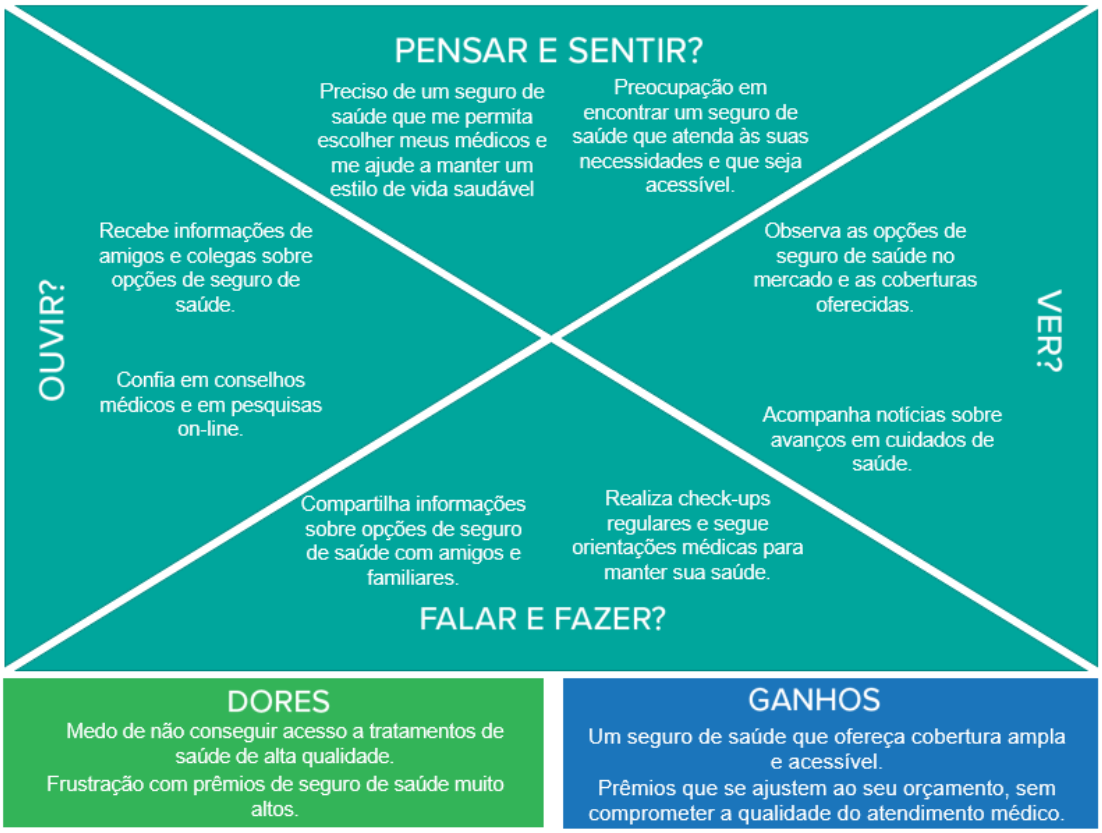
Profissão: Engenheira de Software

Localização: São Paulo, Brasil

Características Comportamentais:

- Preocupa-se com sua saúde e bem-estar.
- Valoriza a liberdade de escolher seus médicos e tratamentos.
- Procura seguros de saúde com prêmios acessíveis.
- Interessada em práticas de sono saudável.

Mapa de Empatia de Ana Silva:



Essas pessoas representam dois grupos de interesse distintos em nosso projeto. Ao compreender suas necessidades, preocupações e comportamentos, estamos melhor preparados para desenvolver estratégias que atendam às expectativas de nossos usuários e da SAFE Seguros.

1.1.3 Benefícios e Justificativas

A implantação de nosso projeto na SAFE Seguros é fundamentada em uma compreensão profunda do contexto do desafio e nas necessidades identificadas em nossas personas, Ana Silva e o Sr. Gilberto Safe.

A seguir, apresentamos os principais fatores que justificam a execução deste projeto e os benefícios futuros esperados:

1. Melhoria na Experiência do Cliente

Com a criação de um modelo de classificação com base nas métricas de sono e outros fatores de estilo de vida, podemos **oferecer aos segurados da SAFE Seguros uma experiência personalizada e de alta qualidade**. Isso inclui ajustar os prêmios de

seguro com base na classificação do tipo de sono, permitindo que os segurados acessem seguros de saúde mais alinhados com suas necessidades e preocupações de saúde.

2. Maior Competitividade no Mercado

A implementação de inovações em nosso portfólio de seguros de saúde nos coloca em uma posição competitiva mais forte. O mercado de seguros de saúde é altamente dinâmico, e a capacidade de **oferecer prêmios personalizados com base em métricas de saúde é um diferencial significativo**. Isso pode atrair novos clientes e manter os atuais segurados satisfeitos com nossos produtos.

3. Redução de Riscos e Custos

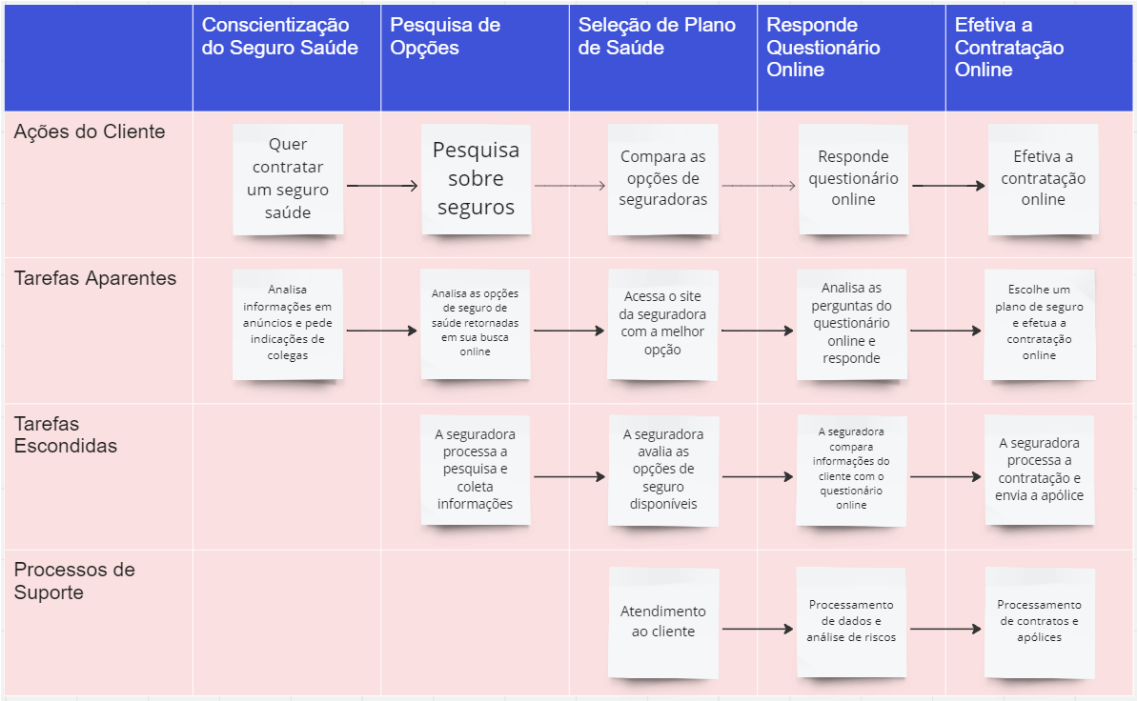
Ao incentivar práticas de saúde preventiva e um estilo de vida saudável, esperamos reduzir os riscos de doenças graves entre nossos segurados. Isso, por sua vez, pode resultar em menores custos de tratamento médico e, potencialmente, prêmios de seguro mais baixos a longo prazo. Além disso, a segmentação de riscos mais precisa nos permite otimizar a alocação de recursos.

4. Impacto Positivo na Sociedade

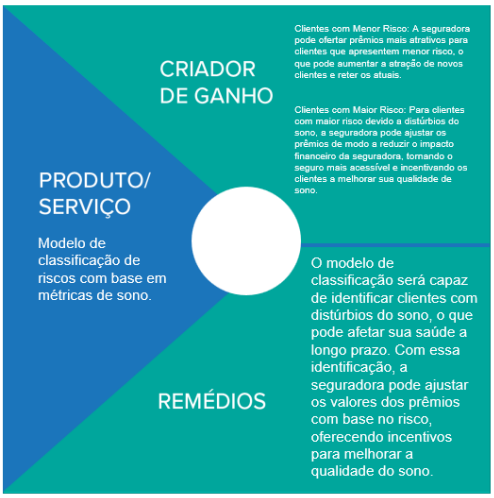
Promover uma abordagem de saúde preventiva tem impactos significativos na sociedade. Nossos segurados terão a oportunidade de melhorar sua qualidade de vida por meio de práticas de sono saudáveis e cuidados de saúde proativos. Isso não apenas beneficia nossos clientes, mas também contribui para a promoção da saúde em nossa comunidade.

Para mapear e comunicar com eficácia a proposta de valor de nosso projeto, utilizamos as seguintes ferramentas estratégicas:

Blueprint: Esta ferramenta nos ajuda a visualizar a jornada do cliente, identificando pontos de contato e interações ao longo do processo. Isso nos permite otimizar a experiência do cliente e aprimorar a eficácia de nossas soluções.



Canvas de Proposta de Valor: Utilizamos o Canvas de Proposta de Valor para definir claramente a proposta de valor de nosso projeto, incluindo os segmentos de cliente a serem atendidos e os benefícios-chave oferecidos.



O investimento em nosso projeto é justificado pelos benefícios significativos que ele trará para nossos segurados, para a empresa e para a sociedade como um todo. A combinação de uma experiência do cliente aprimorada, maior competitividade, redução de custos e impacto social positivo solidifica nossa visão de melhorar a qualidade de vida de nossos clientes e fortalecer a posição da SAFE Seguros no mercado de seguros de saúde.

1.1.4 Hipóteses

Este capítulo apresenta as hipóteses que irão direcionar o desenvolvimento da solução, com base no conhecimento aprofundado do contexto do desafio e nas definições das personas. Antes de prosseguir com o desenvolvimento da solução, é fundamental validar o problema e ter hipóteses sólidas para orientar o projeto.

A matriz de observações para hipóteses é uma ferramenta valiosa para transformar as observações feitas nas etapas anteriores em hipóteses direcionadoras. Ela nos ajuda a entender por que determinadas situações ocorrem e a buscar respostas para essas questões. Abaixo estão algumas das observações transformadas em hipóteses:

Observação	Hipótese
Pessoas frequentemente relatam ter sono ruim devido ao estresse.	Hipótese 1: O nível de estresse influencia na qualidade de sono
Estudos anteriores mostram uma correlação entre a prática regular de atividade física e uma melhor qualidade de sono	Hipótese 2: Atividade física melhora a qualidade de sono.
Há evidências de que a idade está relacionada a mudanças nos padrões de sono, como insônia e fragmentação do sono.	Hipótese 3: A idade pode ser relevante na qualidade do sono.
Profissões com alto nível de estresse ou horários irregulares podem impactar a duração e qualidade do sono dos indivíduos.	Hipótese 4: A profissão pode influenciar na duração e qualidade do sono.

1.2 Solução

1.2.1 Objetivo SMART

O objetivo SMART deste projeto é desenvolver e implementar um modelo de classificação de riscos que utilize métricas de sono para ajustar prêmios de seguros de saúde, permitindo a personalização dos valores com base na qualidade do sono dos clientes. O sucesso será medido pelo aumento da adesão a seguros de saúde personalizados, pela redução de sinistros relacionados a distúrbios do sono e será alcançado até o final do próximo trimestre.

1.2.2 Premissas e Restrições

Neste capítulo, identificamos as premissas e restrições fundamentais que orientarão o desenvolvimento do projeto.

Premissas:

1. **Disponibilidade de Dados de Monitoramento do Sono**: A premissa essencial para o sucesso deste projeto é que haverá acesso aos dados de sono dos clientes por meio de dispositivos de monitoramento ou aplicativos de registro. Esses dados são fundamentais para coletar métricas precisas de sono.
2. **Aceitação do Compartilhamento de Dados pelos Clientes:** É pressuposto que os clientes estarão dispostos a compartilhar suas informações de sono com a seguradora. Isso é fundamental para a personalização dos seguros com base na qualidade do sono.
3. **Competência Técnica da Equipe de Desenvolvimento:** O projeto assume que a equipe de desenvolvimento tem a capacidade técnica necessária para criar o modelo de classificação de riscos com base nas métricas de sono.

Consequências das Premissas não Satisfeitas:

- Se não houver acesso a dados confiáveis de sono, o projeto enfrentará obstáculos na implementação do modelo de classificação de riscos.

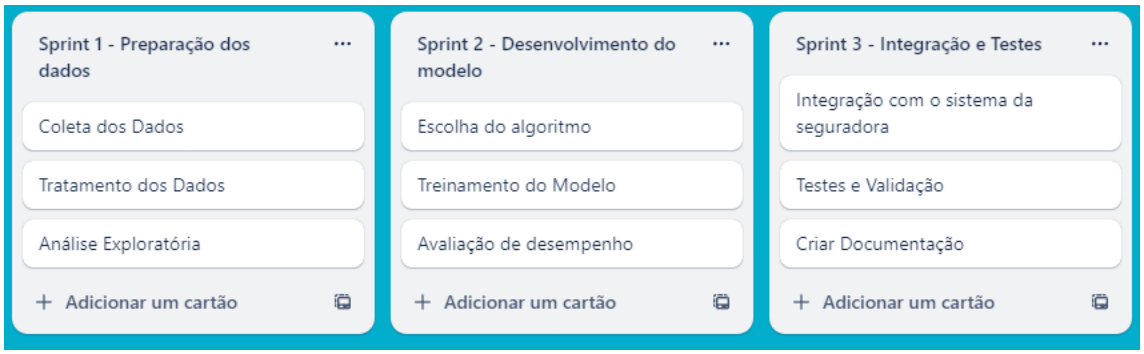
- Se os clientes não aceitarem compartilhar suas informações de sono, a personalização de seguros com base na qualidade do sono será inviável, impactando negativamente a proposta de valor da seguradora.
- A falta de competência técnica na equipe de desenvolvimento pode resultar em atrasos no desenvolvimento do modelo e potencialmente afetar o prazo do projeto.

Restrições:

1. **Orçamento Limitado:** O projeto está sujeito a restrições orçamentárias que podem influenciar a alocação de recursos para o desenvolvimento e implementação. É importante gerenciar eficazmente os recursos disponíveis.
2. **Conformidade com Regulamentações do Setor de Seguros**: O projeto deve aderir estritamente a todas as regulamentações e normas do setor de seguros, o que pode impor limitações às estratégias de precificação com base na qualidade do sono.

1.2.3 Backlog de Produto

O Backlog é uma representação das atividades a serem realizadas no projeto. Com foco no desenvolvimento do Modelo de Classificação, dividimos as atividades em três sprints conforme abaixo:



No próximo capítulo explicaremos detalhadamente cada etapa e começaremos o seu desenvolvimento.

2. Área de Experimentação

Chegamos à etapa de experimentação e desenvolvimento do nosso projeto. Conforme explicitado no capítulo anterior, vamos explicar cada Sprint(etapa) de nossa solução:

Sprint 1 - Preparação de Dados:

- **Coleta de Dados**: Nesta fase, nosso primeiro objetivo é reunir dados disponibilizados pela empresa;
- **Tratamento dos Dados**: Após a coleta, dedicaremos esforços para limpar os dados, eliminando inconsistências;
- **Análise Exploratória**: Analisaremos os dados coletados e tratados para encontrar padrões e insights.

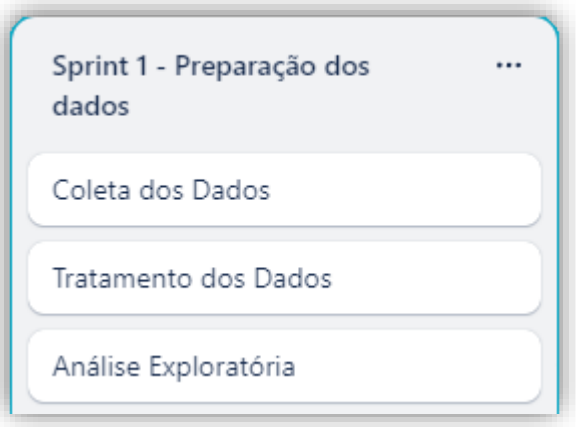
Sprint 2 - Desenvolvimento do Modelo:

- **Escolha do Algoritmo**: Aqui, selecionaremos o algoritmo de classificação mais apropriado para o nosso projeto, com base nas características dos dados.
- **Treinamento do Modelo**: Usando os dados preparados na Sprint 1, treinaremos o modelo de classificação. Isso envolve o ajuste dos parâmetros do algoritmo e a otimização do desempenho.
- **Avaliação de Desempenho**: Avaliaremos o desempenho do modelo com diversas métricas de qualidade, garantindo que ele atenda aos nossos critérios de sucesso.

Sprint 3 - Integração e Testes:

- **Integração com o Sistema da Seguradora**: Após o desenvolvimento, salvaremos o modelo para que seja integrado no sistema da seguradora.
- **Testes e Validação:** Realizaremos testes extensivos para validar o modelo. Isso inclui testes de usabilidade, desempenho e confiabilidade, assegurando que o modelo esteja pronto para produção.
- **Criar Documentação**: Documentaremos o modelo facilitando sua implementação pela seguradora.

2.1 Sprint 1 – Preparação dos Dados



Neste primeiro Sprint nosso foco é a coleta e entendimento dos dados disponibilizados pelo Sr. Gilberto Safe.

2.1.1 Solução

COLETA DOS DADOS

- Evidência do planejamento:

  Nesta etapa o Sr. Gilberto Safe encaminhou o estudo que recebeu de uma clínica do sono por e-mail:

De: clinica_sonhos_perfeitos@gmail.com
Para: Sr. Gilberto Safe ceo@safe-seguros.com
Assunto: Dados de Estudo do Sono para Análise

Prezado, Sr. Gilberto Safe.

Espero que esta mensagem o encontre bem. Em nome da Clínica de Sono "Sonhos Perfeitos", é com grande satisfação que compartilho os dados do estudo de sono que mencionamos durante nosso recente encontro no Congresso de Saúde.

Aqui está o arquivo CSV com os dados do estudo, intitulado "Estudo_Sono_Segurados_SAFE.csv". Este estudo abrangeu um grupo de 400 voluntários e coletou informações detalhadas sobre seus padrões de sono e qualidade do descanso durante um período de seis meses. Os dados foram registrados com o mais alto nível de precisão para garantir a confiabilidade das métricas.

O arquivo contém as seguintes variáveis:

ID do Segurado: Um número de identificação exclusivo para cada voluntário.
Idade: A idade de cada participante no início do estudo.
Sexo: O gênero do voluntário (Masculino/Feminino).
Profissão: A ocupação ou profissão do segurado.
Horas de Sono: O tempo médio de sono (em horas) por noite.
Qualidade do Sono: Uma pontuação subjetiva de 0 a 10, indicando a qualidade percebida do sono.
Nível de atividade física: Uma categoria que reflete o nível de atividade física.
Nível de Stress: Uma categoria que reflete o nível de estresse do voluntário
Índice de Massa Corporal (IMC): Uma medida da composição corporal.
Pressão Arterial
Frequência cardíaca
Passos diários
Distúrbio do sono identificado.

Estamos confiantes de que essas informações serão de grande utilidade para a SAFE Seguros.

Por favor, sinta-se à vontade para entrar em contato caso tenha alguma dúvida ou precise de mais informações. Estamos ansiosos para colaborar com a SAFE Seguros no avanço desta iniciativa inovadora.

Atenciosamente,

Dr. Lucas Sonhos
Clínica de Sono "Sonhos Perfeitos"
Email: clinica_sonhos_perfeitos@gmail.com

Tendo esses dados em mãos, vamos utilizar o Jupyter notebook e a biblioteca Pandas do Python para transformar o dataset em um dataframe para conseguirmos manipular dos dados.

- Evidência da execução de cada requisito:

```
# Carregar os dados
df_sono = pd.read_csv('Estudo_Sono_Segurados_SAFE.csv')

# Visualizar as primeiras linhas
df_sono.head()
```

	ID	sexo	idade	ocupacao	horas_de_sono	qualidade_do_sono	nivel_atividade_fisica	nivel_stress	categoria_IMC	pressao_sanguinea	frequencia_cardiaca	passos_diarios	disturbio_do_sono
0	1	Masculino	27	Engenheiro de Software	6,1	6	42	6	Sobrepeso	126/83	77	4200	NaN
1	2	Masculino	28	Médico	6,2	6	60	8	Normal	125/80	75	10000	NaN
2	3	Masculino	28	Médico	6,2	6	60	8	Normal	125/80	75	10000	NaN
3	4	Masculino	28	Representante de Vendas	5,9	4	30	8	Obeso	140/90	85	3000	Apneia do Sono
4	5	Masculino	28	Representante de Vendas	5,9	4	30	8	Obeso	140/90	85	3000	Apneia do Sono

Acima utilizamos a biblioteca Pandas para carregar os dados e transformá-los em uma estrutura tabelar (Dataframe) para que possamos iniciar nossa análise.

- Evidência dos resultados:

```
# Visualizar o formato(shape) dos dados:
print(f'Número de registros: {df_sono.shape[0]}\nNúmero de variáveis: {df_sono.shape[1]}')
```

```
Número de registros: 374
Número de variáveis: 13
```

Explicando as variáveis dos dados:

Variável	Descrição
ID	Um número de identificação exclusivo para cada voluntário.
sexo	O gênero do voluntário (Masculino/Feminino).
idade	A idade de cada participante no início do estudo.
ocupação	A ocupação ou profissão do segurado.
horas_de_sono	O tempo médio de sono (em horas) por noite.
qualidade_do_sono	Uma pontuação subjetiva de 0 a 10, indicando a qualidade percebida do sono.
nivel_atividade_fisica	Uma categoria que reflete o nível de atividade física.
nivel_stress	Uma classificação subjetiva em uma escala de 1 a 10.
categoria_IMC	Uma medida da composição corporal.
pressao_sanguinea	Indicada como pressão sistólica sobre pressão diastólica.
frequencia_cardiaca	Em batimentos por minuto.
passos_diários	Número de passos diários do voluntário.
disturbio_do_sono	Um dos seguintes - Nenhum, Insônia ou Apneia do Sono.

Como resultado temos um dataframe com 374 registros e 13 variáveis explicadas acima, possibilitando o início da exploração dos dados.

TRATAMENTO DOS DADOS

- Evidência do planejamento:

Com os dados disponibilizados e carregados em um dataframe, vamos começar a verificação.

Verificaremos se as variáveis estão no tipo correto (texto, numéricos, categóricos), se existem dados nulos ou ausentes, e outros ajustes.

- Evidência da execução de cada requisito:

Verificando as informações dos dados:

```
# Vendo informações e formato dos dados:
df_sono.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     374 non-null    int64
 1   sexo                   374 non-null    object
 2   idade                  374 non-null    int64
 3   ocupacao               374 non-null    object
 4   horas_de_sono          374 non-null    object
 5   qualidade_do_sono      374 non-null    int64
 6   nivel_atividade_fisica 374 non-null    int64
 7   nivel_stress           374 non-null    int64
 8   categoria_IMC          374 non-null    object
 9   pressao_sanguinea      374 non-null    object
 10  frequencia_cardiaca    374 non-null    int64
 11  passos_diarios         374 non-null    int64
 12  disturbio_do_sono      155 non-null    object
dtypes: int64(7), object(6)
memory usage: 38.1+ KB
```

Pontos de atenção:

- horas_de_sono: esta com dtype incorreto;
- disturbio_do_sono: Possui dados ausentes ou nulos.

Corrigindo a variável “horas_de_sono”:

Para a variável horas_de_sono, os decimais estão separados por vírgula, por isso foi identificado como object. Abaixo vamos efetuar a correção:

```
# Vamos substituir a ',' por '.' e em seguida alterar o tipo para float
df_sono['horas_de_sono'] = df_sono['horas_de_sono'].str.replace(',','.').astype('float')
```

```
# Confirmando a alteração:
df_sono['horas_de_sono'].dtype
```

```
dtype('float64')
```

Verificando dados ausentes:

Verificando dados ausentes:

```
print(f'Número de registros com dados ausentes:\n{df_sono.isna().sum()}')
```

```
Número de registros com dados ausentes:
ID                          0
sexo                        0
idade                       0
ocupacao                    0
horas_de_sono               0
qualidade_do_sono           0
nivel_atividade_fisica      0
nivel_stress                0
categoria_IMC               0
pressao_sanguinea           0
frequencia_cardiaca         0
passos_diarios              0
disturbio_do_sono         219
```

```
# Verificando os valores contidos na variável:
df_sono['disturbio_do_sono'].value_counts()
```

```
disturbio_do_sono
Apneia do Sono    78
Insônia           77
Name: count, dtype: int64
```

Conforme descrito no início, essa variável deve ter três distúrbios (Nenhum, Apneia, Insônia). Ao carregar os dados, a categoria "Nenhum" foi assumida como um dado ausente. Abaixo segue a correção:

```
# Corrigindo os dados ausentes da coluna disturbio do sono:
df_sono['disturbio_do_sono'].fillna('Nenhum', inplace=True)
```

```
# Verificando a correção:
df_sono['disturbio_do_sono'].value_counts()
```

```
disturbio_do_sono
Nenhum            219
Apneia do Sono     78
Insônia            77
Name: count, dtype: int64
```

Criando uma variável

Para auxiliar na análise exploratória, acredito que seja importante a criação de uma variável informando se o voluntário apresentou algum distúrbio relacionado ao sono.

```
df_sono['tem_disturbio_do_sono'] = df_sono['disturbio_do_sono'].isin(['Apneia do Sono', 'Insônia'])
```

- Evidência dos resultados:

  Nesta etapa verificamos os dados carregados para descobrir ajustes e correções a serem feitas. Vimos acima que a variável “horas_de_sono” foi carregada com o dtypes object (texto), sendo o correto o formato numérico. Também

verificamos que na variável “distúrbio_do_sono” a categoria nenhum foi carregado como dado ausente/nulo.

Para ambas as variáveis foram aplicadas as correções necessárias para que assim possamos prosseguir com a nossa análise.

ANÁLISE EXPLORATÓRIA

- Evidência do planejamento:

  Chegou o momento de explorarmos os dados carregados e tratados. Nesta etapa vamos visualizar algumas estatísticas sobre os dados e formular perguntas para entender melhor os dados e tentar encontrar insights.

- Evidência da execução de cada requisito:

Estatísticas sobre os dados:

Verificando as estatísticas descritivas das variáveis numéricas

```
# Visualizar estatística descritivas dos dados
df_sono.describe()
```

	ID	idade	horas_de_sono	qualidade_do_sono	nivel_atividade_fisica	nivel_stress	frequencia_cardiaca	passos_diarios
count	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000	374.000000
mean	187.500000	42.184492	7.132086	7.312834	59.171123	5.385027	70.165775	6816.844920
std	108.108742	8.673133	0.795657	1.196956	20.830804	1.774526	4.135676	1617.915679
min	1.000000	27.000000	5.800000	4.000000	30.000000	3.000000	65.000000	3000.000000
25%	94.250000	35.250000	6.400000	6.000000	45.000000	4.000000	68.000000	5600.000000
50%	187.500000	43.000000	7.200000	7.000000	60.000000	5.000000	70.000000	7000.000000
75%	280.750000	50.000000	7.800000	8.000000	75.000000	7.000000	72.000000	8000.000000
max	374.000000	59.000000	8.500000	9.000000	90.000000	8.000000	86.000000	10000.000000

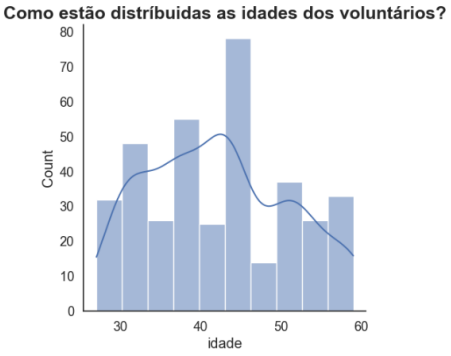
Verificando descrições de dados não numéricos

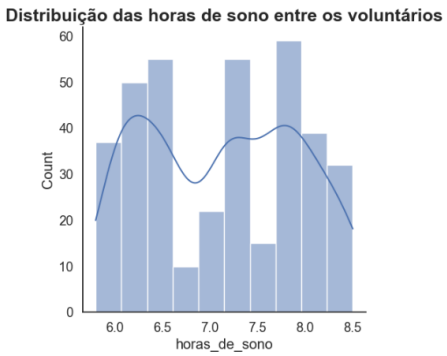
```
df_sono.describe(include=['object','bool'])
```

	sexo	ocupacao	categoria_IMC	pressao_sanguinea	disturbio_do_sono	tem_disturbio_do_sono
count	374	374	374	374	374	374
unique	2	11	4	25	3	2
top	Masculino	Enfermeiro(a)	Normal	130/85	Nenhum	False
freq	189	73	195	99	219	219

Distribuição da Idade e Horas de Sono

Analisando as estatísticas descritivas de todas as variáveis do conjunto de dados, não encontramos nenhum valor discrepante.

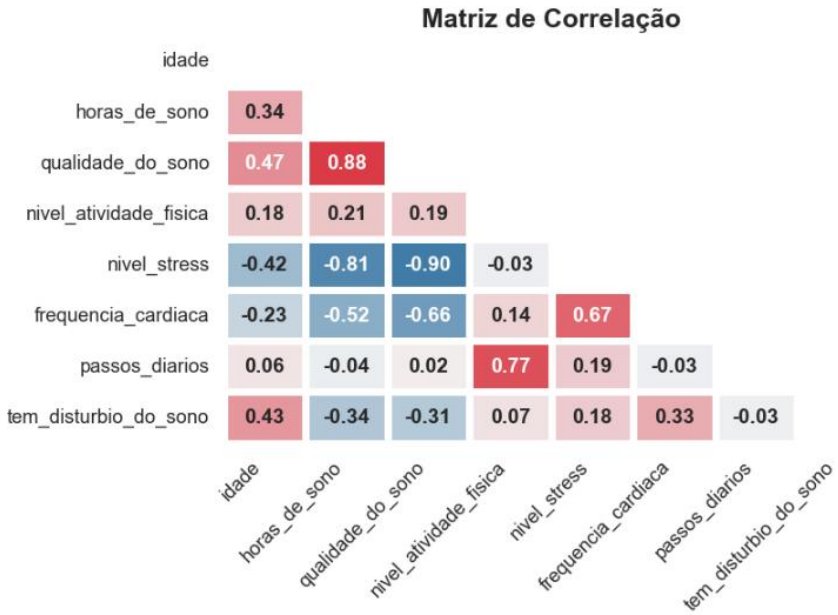




Os histogramas de idade e horas de sono dos voluntários não nos forneceu algum padrão.

Correlação das variáveis

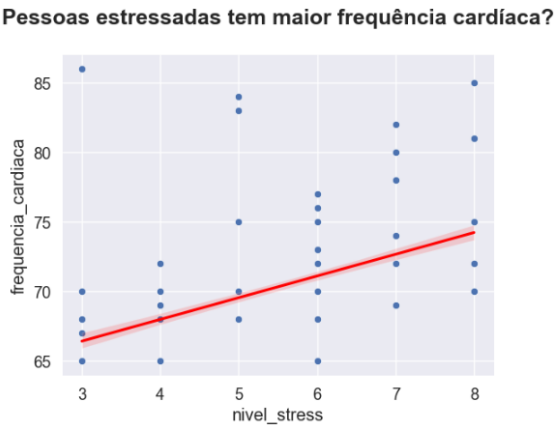
Uma informação relevante a descobrir é: Como as variáveis se relacionam entre si. Para conseguir essa informação vamos criar um mapa de calor e mostrar a correlação entre as variáveis.

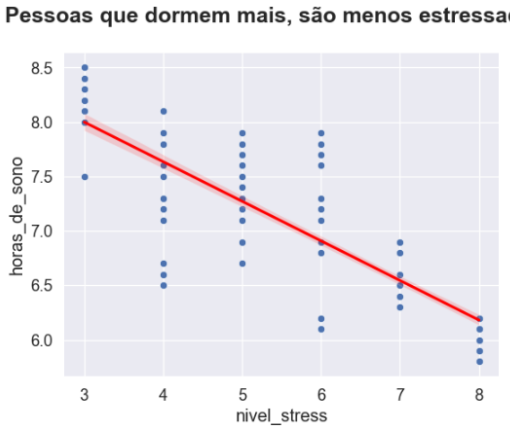


A Matriz de Correlação impressa acima nos trouxe informações interessantes:

- Quem tem mais horas de sono, tem uma qualidade de sono melhor;
- Pessoas com mais nível de stress, também tem maior frequência cardíaca;
- Pessoas com mais horas e com melhor qualidade de sono, são menos estressadas.

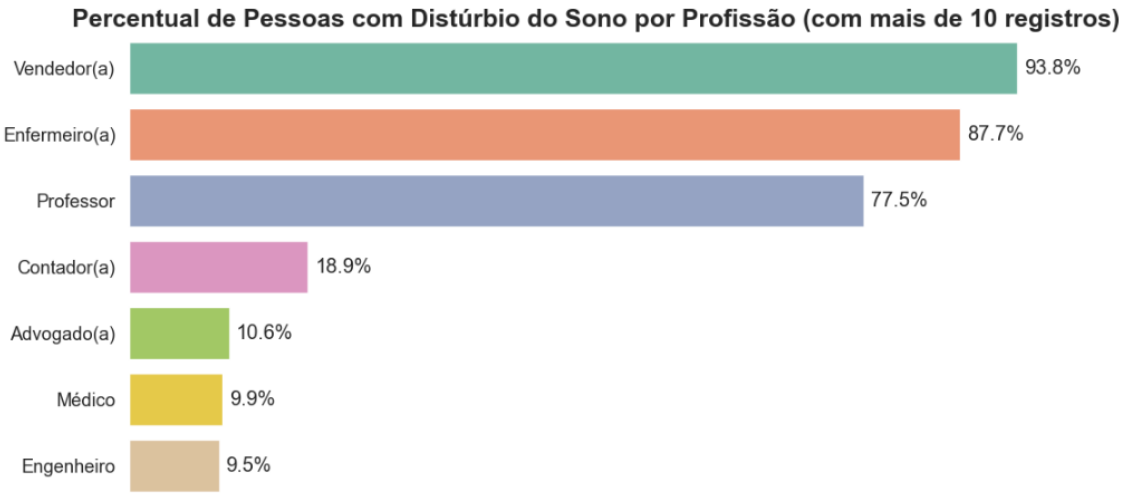
Para evidenciar essas relações, plotamos os gráficos abaixo:

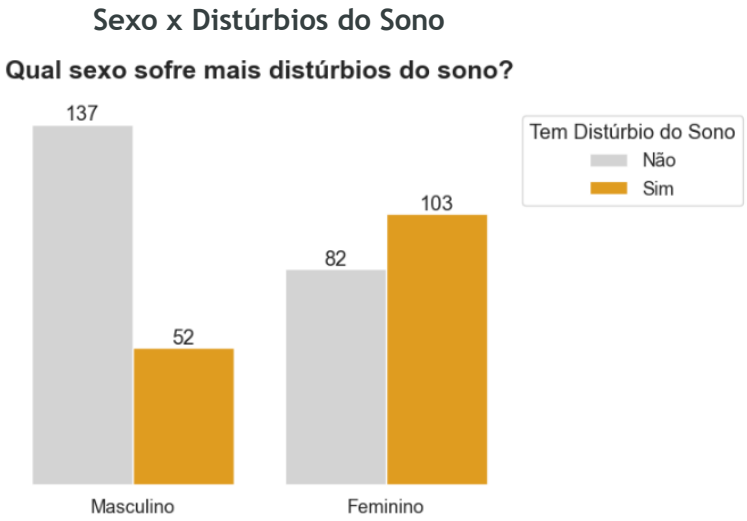




Profissão x Distúrbios do Sono

Outra pergunta informação a ser investigada é qual o profissional que possuí algum distúrbio do sono. Para esta questão, resolvi utilizar apenas as profissões com mais de 10 registro no conjunto de dados:





- Evidência dos resultados:

  Com base na análise exploratória, podemos destacar as seguintes conclusões:

  - A análise revelou que "Vendedor" é a profissão com o maior número de registros de distúrbios do sono.
  - Com base na análise, observamos que as mulheres apresentam mais distúrbios do sono em comparação aos homens. Isso sugere que o sexo pode ser um fator relevante na ocorrência de distúrbios do sono.

2.1.2 Lições Aprendidas

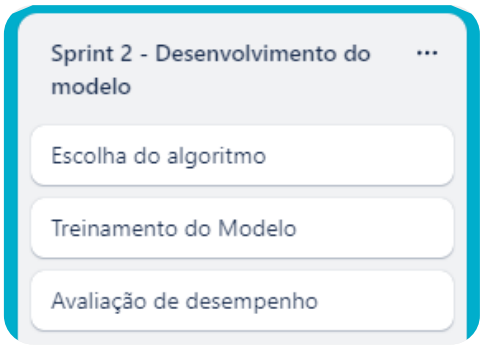
Neste primeiro Sprint destaco a importância de se fazer uma análise inicial nos dados. Vimos que foi necessário correção de variáveis como *horas_de_sono e "distúrbio_do_sono"* afim de garantir uma análise correta dos dados.

Após a análise exploratória obtivemos informações interessantes:

- Sim! Quem dorme mais, dorme melhor;
- E Sim! Quem dorme mais, dorme melhor e é menos estressado.
- Vendedores, enfermeiros e professores são as profissões que mais apresentaram distúrbios de sono;
- Mulheres apresentaram mais distúrbios do sono do que os homens.

Para ver a análise completa, acesse o link:
https://github.com/marcelobin/Projeto-Aplicado-POS-Graduacao-Ciencia-de-Dados-XPEducacao/blob/main/notebooks/sprint1.ipynb

2.2 Sprint 2



2.2.1 Solução

Nesta Sprint vamos fazer as seguintes atividades:

Escolha do Algoritmo: Aqui, selecionaremos o algoritmo de classificação mais apropriado para o nosso projeto, com base nas características dos dados.

Treinamento do Modelo: Usando os dados preparados na Sprint 1, treinaremos o modelo de classificação. Isso envolve o ajuste dos parâmetros do algoritmo e a otimização do desempenho.

Avaliação de Desempenho: Avaliaremos o desempenho do modelo com diversas métricas de qualidade, garantindo que ele atenda aos nossos critérios de sucesso.

Escolha do Algoritmo

- Evidência do planejamento:

Neste projeto, serão utilizados três algoritmos de classificação para modelar e prever os resultados desejados. Cada um desses algoritmos tem características distintas, oferecendo diferentes vantagens em termos de desempenho.

1. RandomForest

O RandomForest é um modelo baseado em conjunto que combina as previsões de várias árvores de decisão para obter uma previsão mais robusta e geral.

Este modelo é conhecido por sua eficácia em lidar com uma variedade de conjuntos de dados e sua resistência ao sobre ajuste.

2. Logistic Regression

O Logistic Regression é uma técnica de regressão utilizada para problemas de classificação binária. Apesar da sua simplicidade, a Regressão Logística é rápida e fácil de interpretar.

É especialmente eficaz quando a relação entre as features e a variável de destino é predominantemente linear.

3. Decision Tree

O Decision Tree é um modelo de árvore de decisão que divide iterativamente o conjunto de dados em subconjuntos mais puros.

Apesar de sua simplicidade, as árvores de decisão podem ser propensas a sobre ajuste, ajustando-se demais aos dados de treinamento.

- Evidência da execução de cada requisito:

Para a implantação desses algoritmos, utilizamos a biblioteca de Machine Learning do Python chamada Scikit-Learn.

- Evidência dos resultados:

Abaixo a importação da biblioteca e módulos:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, export_graphviz
```

Treinamento do Modelo

- Evidência do planejamento:

Para o treinamento dos modelos, vamos utilizar a função do Scikit-Learn GridSearchCV.

O GridSearchCV encontra os melhores hiper parâmetros para um modelo por meio da busca em uma grade de valores possíveis para esses hiper parâmetros.

Além disso, ele também divide o conjunto de dados em partes (folds) e avalia o desempenho do modelo usando validação cruzada em cada combinação de hiper parâmetros.

- Evidência da execução de cada requisito:

Abaixo segue o código de definição dos hiper parâmetros, criação dos classificadores, treinamento dos modelos e, finalmente, os melhores parâmetros encontrados pelo GridSearchCV

```
# Definindo os hiperparâmetros que deseja otimizar para cada modelo
param_grid_rf = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10, 15, 20],
    'min_samples_leaf': [1, 2, 4, 6, 8]
}

param_grid_lr = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2']
}

param_grid_dt = {
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10, 15, 20],
    'min_samples_leaf': [1, 2, 4, 6, 8]
}

# Criando os classificadores
rf_classifier = RandomForestClassifier(random_state=23)
lr_classifier = LogisticRegression(random_state=23)
dt_classifier = DecisionTreeClassifier(random_state=23)

# Use GridSearchCV para encontrar os melhores hiperparâmetros para cada modelo
grid_search_rf = GridSearchCV(rf_classifier, param_grid_rf, cv=5, scoring='accuracy')
grid_search_lr = GridSearchCV(lr_classifier, param_grid_lr, cv=5, scoring='accuracy')
grid_search_dt = GridSearchCV(dt_classifier, param_grid_dt, cv=5, scoring='accuracy')

# Treinar os modelos com os melhores hiperparâmetros
grid_search_rf.fit(X_train, y_train)
grid_search_lr.fit(X_train, y_train)
grid_search_dt.fit(X_train, y_train)

# Imprimir os melhores hiperparâmetros para cada modelo
print("Melhores hiperparametros para RandomForest:", grid_search_rf.best_params_)
print("Melhores hiperparametros para Logistic Regression:", grid_search_lr.best_params_)
print("Melhores hiperparametros para Decision Tree:", grid_search_dt.best_params_)
```

- Evidência dos resultados:

Abaixo os melhores hiper parâmetros encontrados para cada algoritmo:

```
Melhores hiperparametros para RandomForest: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Melhores hiperparametros para Logistic Regression: {'C': 0.01, 'penalty': 'l2'}
Melhores hiperparametros para Decision Tree: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10}
```

Avaliação de Desempenho

- Evidência do planejamento:

Para avaliação dos algoritmos, vamos utilizar seis métricas:

- **ACURÁCIA:** Mede a proporção de predições corretas em relação ao total de predições.
- **PRECISÃO:** Mede a proporção de instâncias positivas previstas corretamente em relação ao total de instâncias positivas previstas.
- **RECALL:** Mede a proporção de instâncias positivas previstas corretamente em relação ao total de instâncias positivas reais.
- **F1-SCORE:** É a média harmônica entre precisão e recall.
- **AUC-ROC:** É uma métrica que avalia o desempenho do modelo em diferentes pontos de corte de probabilidade. Um valor próximo de 1.0 indica um bom desempenho.
- **MATRIZ DE CONFUSÃO:** A matriz de confusão é uma tabela que mostra a distribuição dos resultados da classificação em cada classe. Fornece uma visão detalhada dos resultados do modelo, incluindo verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

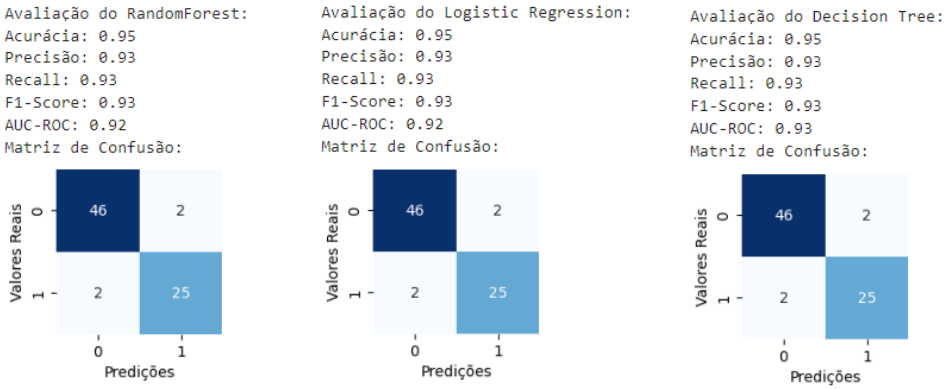
- Evidência da execução de cada requisito:

Para efetuar a avaliação de cada modelo, decidi criar uma função que executa a medição de todas as métricas citadas acima:

```
# Função para avaliar e imprimir métricas
def avalia_modelo(model, X_test, y_test):
    y_pred = model.predict(X_test)

    # Acurácia
    acuracia = accuracy_score(y_test, y_pred)
    print("Acurácia:", round(acuracia,2))

    # Precisão
    precisao = precision_score(y_test, y_pred)
    print("Precisão:", round(precisao,2))

    # Recall
    recall = recall_score(y_test, y_pred)
    print("Recall:", round(recall,2))

    # F1-Score
    f1 = f1_score(y_test, y_pred)
    print("F1-Score:", round(f1,2))

    # AUC-ROC
    if hasattr(model, 'predict_proba'):
        y_probs = model.predict_proba(X_test)[:, 1]
        auc_roc = roc_auc_score(y_test, y_probs)
        print("AUC-ROC:", round(auc_roc,2))

    # Matriz de Confusão
    conf_matrix = confusion_matrix(y_test, y_pred)
    print("Matriz de Confusão:")

    # Criar um mapa de calor usando seaborn
    plt.figure(figsize=(2, 2))
    sns.heatmap(conf_matrix, annot=True, cmap="Blues", cbar=False)
    plt.xlabel('Predições')
    plt.ylabel('Valores Reais')
    plt.show()
```

- Evidência dos resultados:

Após aplicar a função, obtivemos os resultados abaixo:



Todos os modelos tiveram resultados praticamente idênticos.

2.2.2 Lições Aprendidas

Na escolha do modelo para nosso projeto, decidimos apostar no Logistic Regression como nossa ferramenta principal. A decisão baseia-se em resultados consistentes e similarmente sólidos entre RandomForest, Decision Tree e Logistic Regression.

Optamos pelo Logistic Regression devido à sua simplicidade e interpretabilidade, uma escolha que faz sentido para o tamanho de nossa base de dados e facilita a compreensão do funcionamento do modelo.

Link do código:

https://github.com/marcelobin/Projeto-Aplicado-POS-Graduacao-Ciencia-de-Dados-XPEducacao/blob/main/notebooks/Sprint%202%20-%20Desenvolvimento%20do%20Modelo.ipynb

2.3 Sprint 3

2.3.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.3.2 Lições Aprendidas

3. Considerações Finais

3.1 Resultados

Por meio de um texto detalhado, apresente os principais resultados alcançados pelo seu Projeto Aplicado.

Cite os pontos positivos e negativos, as dificuldades enfrentadas e as experiências vivenciadas durante todo o processo.

3.2 Contribuições

Apresente quais foram as contribuições que o seu Projeto Aplicado trouxe para que o Desafio proposto fosse solucionado.

Cite, por exemplo, as inovações, as vantagens sobre os similares, as melhorias alcançadas, entre outros.

3.3 Próximos passos

Descreva quais são os próximos passos que poderão contribuir com o aprimoramento da solução apresentada pelo seu Projeto Aplicado.