UNIVERSIDADE PRESBITERIANA MACKENZIE

CURSO DE GRADUAÇÃO EM CIÊNCIA DE DADOS

PROJETO APLICADO II

ÍNDICE GLOBAL DE DESENVOLVIMENTO E PROSPERIDADE

São Paulo

2024

LISTA DE FIGURAS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

No cenário atual, o acesso a dados de qualidade e a capacidade de analisá-los de forma eficaz são fatores essenciais para a tomada de decisões estratégicas em qualquer organização. Este projeto tem como objetivo desenvolver um estudo prático utilizando os dados do Legatum Institute, uma organização dedicada à pesquisa e análise de indicadores globais de prosperidade. O foco deste estudo será aplicar técnicas de ciência de dados para extrair insights significativos que possam contribuir para uma melhor compreensão dos fatores que influenciam a prosperidade em diferentes países oferecendo uma perspectiva sobre como diferentes aspectos sociais e políticos contribuem para o bem-estar geral de uma nação.

A proposta consiste em explorar os dados disponibilizados pelo Legatum Institute, que incluem uma série de indicadores relacionados à segurança, liberdade pessoal, governança, capital social, ambiente de investimento, condições empresariais, e outros aspectos críticos que afetam o desenvolvimento econômico e social das nações.

2 SOBRE A ORGANIZAÇÃO

O Legatum Institute é uma organização internacional de pesquisa e política pública com sede em Londres, Reino Unido. Fundado em 2007, o instituto tem como missão promover o desenvolvimento de sociedades mais prósperas ao redor do mundo. O Legatum Institute é particularmente conhecido pelo seu trabalho na medição e análise da prosperidade global, sendo o criador do Legatum Prosperity Index, um dos índices mais respeitados que avaliam o bem-estar das nações.

O Legatum Prosperity Index é uma ferramenta abrangente que mede a prosperidade em diversos países com base em uma variedade de fatores, como qualidade econômica, governança, educação, saúde, segurança, liberdade pessoal, capital social, e condições ambientais. O índice tem o objetivo de fornecer uma visão holística da prosperidade, indo além das métricas econômicas tradicionais, e oferecendo uma perspectiva sobre como diferentes aspectos sociais e políticos contribuem para o bem-estar geral de uma nação.

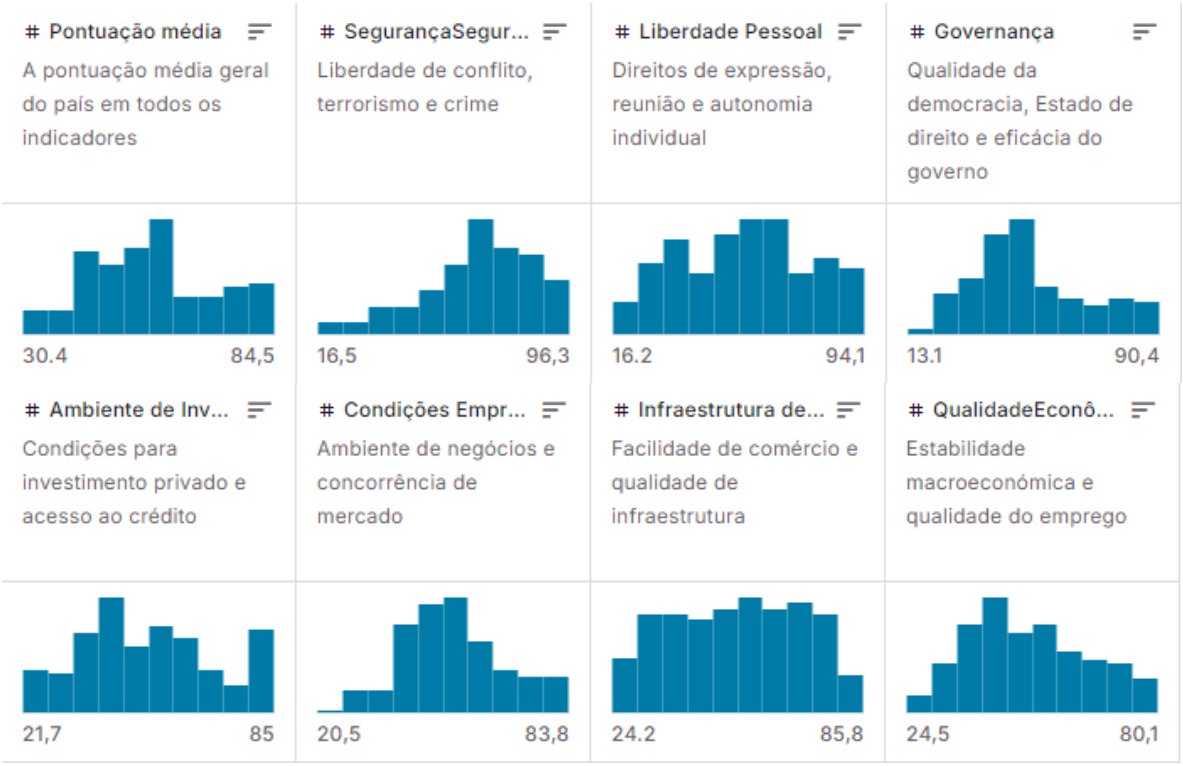
2.1 ÁREA DE ATUAÇÃO

O projeto será desenvolvido na área de Análise de Dados de Prosperidade e Desenvolvimento Global, com foco na investigação dos fatores que influenciam o bem-estar e o desenvolvimento econômico, social e ambiental das nações. Utilizando o dataset do Índice Global de Desenvolvimento e Prosperidade dos Países de 2023, o estudo se concentrará na análise exploratória de dados para identificar padrões, tendências e correlações entre diversos indicadores de prosperidade.

2.1.1 Necessidades e problemas

1. **Avaliar a Prosperidade Global:** Existe a necessidade de realizar uma análise abrangente que explore como os diferentes indicadores contribuem para a prosperidade e de identificar padrões e tendências que possam informar políticas públicas e estratégias de desenvolvimento.
2. **Identificar Fatores Críticos para a Prosperidade:** O objetivo é identificar e priorizar os fatores que têm o maior impacto na prosperidade de um país. Isso permitirá que governos e organizações concentrem seus esforços nas áreas que mais precisam de melhorias.
3. **Compreender Disparidades Regionais e Globais:** É necessário mapear e compreender as disparidades regionais e globais nos indicadores de prosperidade para identificar áreas geográficas ou países que necessitam de intervenções específicas.
4. **Análise da Relação entre Indicadores:** O objetivo é realizar uma análise exploratória para revelar correlações e causalidades entre os diferentes indicadores, ajudando a entender como melhorias em uma área podem influenciar positivamente outras áreas.

Gráfico 1 – Distribuição de diferentes indicadores de prosperidade



3 APRESENTAÇÃO DOS DADOS

Para o desenvolvimento deste projeto, será utilizado um conjunto de dados quantitativos fornecido pelo Legatum Institute, que inclui uma série de indicadores relacionados à prosperidade e ao desenvolvimento global dos países em 2023. Os dados serão apresentados e analisados através de visualizações gráficas. Os indicadores quantitativos e categorias são bem definidos (como país e diferentes dimensões de prosperidade), portanto, gráficos serão a principal ferramenta utilizada para ilustrar as tendências e insights derivados dos dados.

3.1 DADOS UTILIZADOS

O dataset contém os seguintes indicadores, que serão explorados e apresentados:

- **Country:** O nome do país.
- **AveragScore:** A pontuação média geral do país em todos os indicadores.

- **SafetySecurity:** Liberdade de conflito, terrorismo e crime.
- **PersonelFreedom:** Direitos de expressão, reunião e autonomia individual.
- **Governance:** Qualidade da democracia, estado de direito e eficácia do governo.
- **SocialCapital:** Força dos relacionamentos pessoais e engajamento cívico.
- **InvestmentEnvironment:** Condições para investimento privado e acesso ao crédito.
- **EnterpriseConditions:** Ambiente de negócios e concorrência de mercado.
- **MarketAccessInfrastructure:** Facilidade de comércio e qualidade de infraestrutura.
- **EconomicQuality:** Estabilidade macroeconômica e qualidade do emprego.
- **LivingConditions:** Padrão de vida e acesso a serviços básicos.
- **Health:** Saúde da população e acesso à assistência médica.
- **Education:** Qualidade e acessibilidade da educação.
- **NaturalEnvironment:** Qualidade ambiental e sustentabilidade.

Para realizar uma análise exploratória nestes dados, alguns passos serão seguidos:

1. Limpeza e preparação dos dados:

- Tratar valores ausentes, dados duplicados e outliers;

2. Exploração Inicial dos dados:

- Visualizar as primeiras linhas dos dados para entender a estrutura;
- Utilizar funções como ‘summary()’ para obter uma visão geral estatística dos dados.

3. Estatística:

- Resumo Estatístico: Calcular estatísticas descritivas para cada indicador, como média, mediana, e desvio padrão, ajudando a entender a distribuição de cada variável.

Figura 1 – Índice Global de Desenvolvimento e Prosperidade dos Países de 2023

	Country	AveragScore	SafetySecurity	PersonelFreedom	Governance	\
0	Denmark	84.55	92.59	94.09	89.45	
1	Sweden	83.67	90.97	91.90	86.41	
2	Norway	83.59	93.30	94.10	89.66	
3	Finland	83.47	89.56	91.96	90.41	
4	Switzerland	83.42	95.66	87.50	87.67	

	SocialCapital	InvestmentEnvironment	EnterpriseConditions	\
0	82.56	82.42	79.64	
1	78.29	82.81	75.54	
2	79.03	82.24	75.95	
3	77.27	84.12	77.25	
4	69.14	80.81	83.84	

	MarketAccessInfrastructure	EconomicQuality	LivingConditions	Health	\
0	78.79	76.81	95.77	81.07	
1	79.67	76.18	95.33	82.28	
2	75.87	77.25	94.70	82.98	
3	78.77	70.28	94.46	81.19	
4	78.65	79.71	94.66	82.11	

	Education	NaturalEnvironment
0	87.48	73.94
1	85.92	78.74
2	85.68	72.37
3	88.38	77.99
4	87.72	73.60

3.2 ANÁLISE EXPLORATÓRIA

Nesta seção será apresentada uma análise exploratória dos dados do Índice Global de Desenvolvimento e Prosperidade dos Países de 2023, a análise será realizada para examinar os indicadores apresentados anteriormente.

Esta análise identificará padrões, tendências, possíveis anomalias e correlações entre as variáveis, preparando para etapas mais avançadas de modelagem e tomada de decisões. Além disso, a análise auxiliará no tratamento de dados ausentes e na normalização das variáveis, garantindo que o conjunto de dados esteja pronto para uso em modelos analíticos.

- **Análise Inicial**

Figura 2 - Número de colunas e tipos de dados

```
Data columns (total 14 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Country                     167 non-null    object
 1   AveragScore                 167 non-null    float64
 2   SafetySecurity              167 non-null    float64
 3   PersonelFreedom             167 non-null    float64
 4   Governance                  167 non-null    float64
 5   SocialCapital               167 non-null    float64
 6   InvestmentEnvironment       167 non-null    float64
 7   EnterpriseConditions        167 non-null    float64
 8   MarketAccessInfrastructure  167 non-null    float64
 9   EconomicQuality             167 non-null    float64
 10  LivingConditions            167 non-null    float64
 11  Health                      167 non-null    float64
 12  Education                   167 non-null    float64
 13  NaturalEnvironment          167 non-null    float64
dtypes: float64(13), object(1)
memory usage: 18.4+ KB
```

Figura 3 - Contagem de valores, média, desvio padrão, valor mínimo, quartis e valor máximo

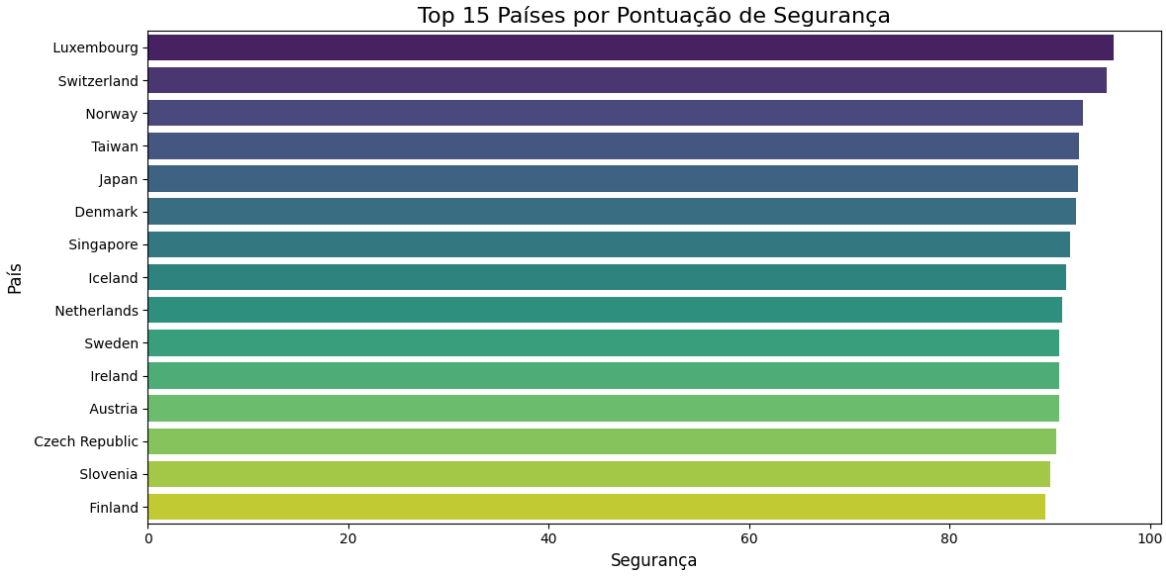
	AveragScore	SafetySecurity	PersonelFreedom	Governance	SocialCapital	InvestmentEnvironment	EnterpriseConditions
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	58.056228	67.242515	56.829401	50.360299	54.468024	53.013653	54.791677
std	13.309964	17.542279	19.914638	17.747092	10.350639	16.743723	12.876166
min	30.400000	16.540000	16.160000	13.090000	23.010000	21.690000	20.500000
25%	47.770000	59.290000	39.650000	37.470000	47.820000	40.765000	45.830000
50%	57.530000	68.930000	57.170000	47.510000	54.390000	51.250000	53.520000
75%	66.860000	80.560000	72.865000	60.965000	60.610000	64.640000	62.760000
max	84.550000	96.320000	94.100000	90.410000	82.560000	84.990000	83.840000

MarketAccessInfrastructure	EconomicQuality	LivingConditions	Health	Education	NaturalEnvironment
167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
55.019701	51.568683	69.803293	68.624910	58.723772	56.231737
15.849004	13.427811	19.752677	11.166266	19.679023	9.061262
24.230000	24.460000	19.210000	31.950000	16.780000	33.670000
40.055000	41.295000	55.830000	60.595000	44.355000	50.265000
56.590000	50.060000	74.770000	71.380000	61.930000	55.540000
69.060000	62.555000	86.975000	77.340000	74.130000	61.940000
85.750000	80.100000	95.860000	86.890000	91.440000	78.740000

Figura 4 - Verificação de valores ausentes

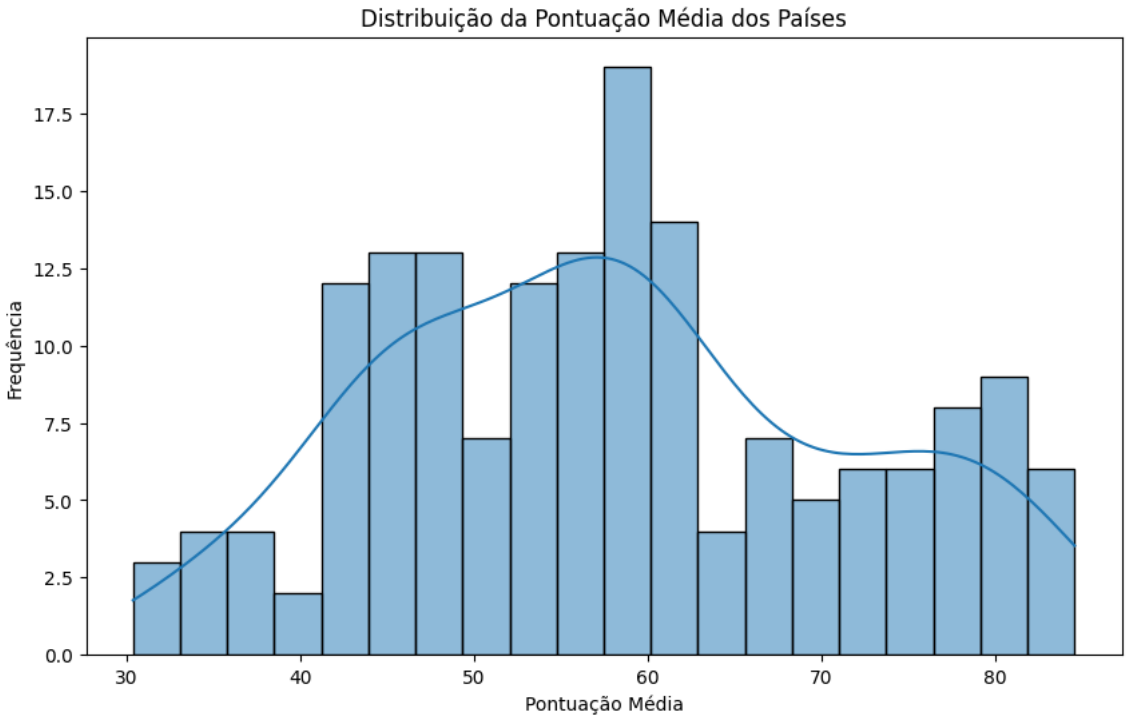
	0		
Country	0.0	EnterpriseConditions	0.0
AveragScore	0.0	MarketAccessInfrastructure	0.0
SafetySecurity	0.0	EconomicQuality	0.0
PersonelFreedom	0.0	LivingConditions	0.0
Governance	0.0	Health	0.0
SocialCapital	0.0	Education	0.0
InvestmentEnvironment	0.0	NaturalEnvironment	0.0

Gráfico 2 - Análise Comparativa (Top 15 Países por Pontuação de Segurança)



A análise inicial foi essencial para fornecer uma visão clara da estrutura básica dos dados e das principais colunas. Identificamos estatísticas descritivas, valores ausentes e visualizamos a distribuição dos dados através de gráficos de barras, que comparou de forma simples as pontuações de segurança entre países, e também a distribuição da pontuação média dos países, como mostrado a seguir:

Gráfico 3 - Análise de Distribuição de Variáveis Numéricas



Além de proporcionar um entendimento preliminar dos dados, essa análise identificou possíveis outliers e inconsistências, o que ajudará a direcionar as próximas etapas do tratamento. Com isso, será possível aplicar técnicas adequadas e evitar erros ao tratar os dados, garantindo uma preparação mais precisa para análises futuras.

- **Tratamento da Base de Dados**

Após a análise exploratória inicial, é necessário preparar os dados para remover ruídos, corrigir inconsistências, lidar com valores ausentes e normalizar as variáveis, assegurando uma melhor performance nos modelos subsequentes. Esta etapa também envolve a codificação de variáveis categóricas e a divisão dos dados em conjuntos de treino e teste. Com isso, os dados se tornam mais consistentes e adequados para fornecer resultados precisos em análises preditivas ou descritivas.

1. Normalização

A normalização é uma técnica usada em ciência de dados para padronizar os valores de diferentes variáveis, colocando-os dentro de uma faixa comum. No caso deste exemplo, utilizamos o MinMaxScaler da biblioteca sklearn, que transforma os valores de uma variável para um intervalo entre 0 e 1, ou outro intervalo definido. Essa abordagem é útil especialmente quando estamos lidando com variáveis em escalas diferentes, pois facilita a comparação e o uso em algoritmos de aprendizado de máquina que são sensíveis à escala dos dados.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

dados[['SafetySecurity', 'PersonelFreedom', 'Governance']] = scaler.fit_transform(
    dados[['SafetySecurity', 'PersonelFreedom', 'Governance']]
)
```

2. Codificação de Variáveis Categóricas

A codificação de variáveis categóricas é uma técnica usada para converter dados que estão em formato de texto (ou seja, categorias) para valores numéricos. Muitos algoritmos de aprendizado de máquina não conseguem lidar diretamente com dados categóricos, então essa transformação é essencial. No exemplo abaixo, usamos a classe LabelEncoder da biblioteca sklearn para transformar a coluna Country (países) em valores numéricos.

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()

dados['Country_encoded'] = label_encoder.fit_transform(dados['Country'])

print(dados[['Country', 'Country_encoded']].head())
```

3. Divisão dos Dados

A divisão dos dados em conjuntos de treino e teste é uma etapa fundamental no desenvolvimento de modelos de aprendizado de máquina. Ela garante que o modelo seja avaliado em dados que ele ainda não viu, permitindo uma medição mais precisa de sua performance. No exemplo a seguir, usamos a função train_test_split da biblioteca sklearn.model_selection para dividir os dados.

```
from sklearn.model_selection import train_test_split

X = dados.drop(columns=['AveragScore'])
y = dados['AveragScore']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Durante o processo de tratamento da base de dados, diversas técnicas essenciais foram aplicadas para garantir que o modelo seja treinado em um conjunto de dados limpo, estruturado e preparado, minimizando a presença de inconsistências que poderiam afetar negativamente o desempenho. Esse tratamento é uma etapa indispensável, pois contribui diretamente para a melhoria da qualidade das previsões e da capacidade de generalização do modelo.

- **Resultados do Tratamento**

A seguir, apresentamos os resultados obtidos de cada uma das etapas do tratamento de dados:

1. Normalização
2. Codificação
3. Divisão de Dados

Figura 5 – Antes e Depois da Normalização

```
Valores Mínimos e Máximos Antes da Normalização:
SafetySecurity     16.54
PersonelFreedom    16.16
Governance         13.09
dtype: float64
SafetySecurity     96.32
PersonelFreedom    94.10
Governance         90.41
dtype: float64

Valores Mínimos e Máximos Depois da Normalização:
SafetySecurity     0.0
PersonelFreedom    0.0
Governance         0.0
dtype: float64
SafetySecurity     1.0
PersonelFreedom    1.0
Governance         1.0
dtype: float64

Valores Normalizados (Depois da Normalização):
   SafetySecurity  PersonelFreedom  Governance
0        0.953246         0.999872    0.987584
1        0.932941         0.971773    0.948267
2        0.962146         1.000000    0.990300
3        0.915267         0.972543    1.000000
4        0.991727         0.915319    0.964563
```

Figura 6 – Coluna “Country” Codificada

	Country	Country_encoded
0	Denmark	40
1	Sweden	141
2	Norway	110
3	Finland	50
4	Switzerland	165

Figura 7 – Dados divididos em Treino e Teste

```
Tamanho do conjunto de treino: 133
Tamanho do conjunto de teste: 34
```

4 MÉTODOS ANALÍTICOS

A etapa de definição dos métodos analíticos é crucial para fundamentar as análises realizadas, ao estabelecer uma base teórica sólida, é possível justificar a escolha das ferramentas estatísticas utilizadas e garantir que os métodos aplicados estejam adequadamente alinhados com os objetivos da pesquisa. Neste contexto, dois métodos analíticos comumente utilizados em estudos preditivos e exploratórios são o Teste de Hipótese e a Análise de Variância.

Estes métodos permitem não apenas verificar suposições sobre características da população com base em amostras, mas também identificar diferenças significativas entre grupos. O entendimento dos princípios teóricos que fundamentam esses métodos garante que as interpretações dos resultados sejam feitas de forma rigorosa e baseada em evidências estatísticas. A seguir, descrevemos as bases teóricas e a aplicação desses métodos no contexto do nosso estudo.

1. Teste de Hipótese

O teste de hipótese é uma técnica estatística utilizada para avaliar se uma afirmação sobre uma característica de uma população é plausível com base em dados amostrais. Ele envolve a formulação de duas hipóteses:

- Hipótese nula (Ho): representa a suposição de que não há efeito ou diferença significativa. Geralmente, busca-se rejeitar a hipótese nula.
- Hipótese alternativa (Ha): é a suposição que contradiz a hipótese nula e representa a afirmação que queremos testar.

O procedimento consiste em calcular uma estatística de teste a partir dos dados e compará-la com um valor crítico, definido por um nível de significância (geralmente 5%, ou 0.05). Se a estatística de teste ultrapassar o valor crítico, rejeitamos a hipótese nula em favor da alternativa.

Elementos do teste de hipótese:

- **Valor-p**: a probabilidade de observarmos um resultado tão extremo quanto o dos dados, sob a suposição de que a hipótese nula é verdadeira.

- **Nível de significância (α)**: a probabilidade máxima de cometer o erro tipo I (rejeitar Ho quando ela é verdadeira).
- **Erro Tipo I e II**: erros associados ao processo de tomada de decisão. O erro tipo I é rejeitar uma hipótese nula verdadeira, e o erro tipo II é aceitar uma hipótese nula falsa.

Esse método é amplamente utilizado para verificar afirmações sobre médias, proporções, e correlações em diferentes contextos. No caso deste estudo, o teste de hipótese pode ser utilizado, por exemplo, para avaliar se há diferenças significativas em determinadas variáveis de interesse entre diferentes grupos ou populações. O exemplo a seguir verifica se a média do indicador 'AveragScore' difere significativamente de uma média hipotética, 50.

```
from scipy import stats

media_hipotetica = 50

t_stat, p_val = stats.ttest_1samp(dados['AveragScore'], media_hipotetica)

print(f"Estatística t: {t_stat}")
print(f"Valor-p: {p_val}")

# Resultado
if p_val < 0.05:
    print("Rejeitamos a hipótese nula: a média difere significativamente de 50.")
else:
    print("Falhamos em rejeitar a hipótese nula: a média não difere significativamente de 50.")
```

```
Estatística t: 7.821915053158361
Valor-p: 5.744227977972093e-13
Rejeitamos a hipótese nula: a média difere significativamente de 50.
```

Esse resultado do teste de hipótese significa que a média do AveragScore é significativamente diferente de 50. A estatística t de aproximadamente 7,82 indica uma diferença substancial, e o valor-p extremamente baixo (5,74e-13) confirma que essa diferença não é atribuível ao acaso, assumindo um nível de significância comum, como 0,05.

2. Análise de Variância

A análise de variância é um método estatístico que tem como objetivo testar diferenças entre as médias de três ou mais grupos. Ela é utilizada para verificar se essas

diferenças são estatisticamente significativas, ou seja, se as variações observadas são reais ou se podem ser atribuídas ao acaso.

A análise de variância parte das seguintes hipóteses:

- **Hipótese nula (H_0)**: todas as médias dos grupos são iguais.
- **Hipótese alternativa (H_1)**: pelo menos uma das médias dos grupos é diferente.

Na análise de variância, comparamos a variância entre os grupos (variabilidade explicada) com a variância dentro dos grupos (variabilidade não explicada). O resultado é uma estatística F, que permite avaliar a razão entre essas variâncias. Se a estatística F for maior que um valor crítico, rejeitamos a hipótese nula.

Principais conceitos na ANOVA:

- Variância entre grupos: mede as diferenças entre as médias dos grupos.
- Variância dentro dos grupos: mede a dispersão dentro de cada grupo.
- Estatística F: relação entre a variância entre grupos e a variância dentro dos grupos. Uma estatística F alta indica que há diferenças significativas entre os grupos.

A análise de variância é uma ferramenta poderosa para comparar múltiplos grupos e é amplamente aplicada em experimentos controlados e estudos com múltiplas variáveis categóricas. Neste estudo, a análise de variância pode ser usada para analisar diferenças entre as variáveis de segurança e liberdade pessoal em diferentes países ou blocos econômicos, por exemplo. O exemplo a seguir calcula a média das colunas “SafetySecurity”, “PersonelFreedom” e “LivingConditions” para cada país, criando uma métrica agregada. Em seguida, foi testado se há uma diferença significativa entre os grupos formados por essa métrica agregada.

```python
from scipy import stats

dados['Medias'] = dados[['SafetySecurity', 'PersonelFreedom', 'LivingConditions']].mean(axis=1)

dados['Grupo'] = pd.qcut(dados['Medias'], q=4, labels=['Baixo', 'Médio-baixo', 'Médio-alto', 'Alto'])


print(dados['Grupo'].value_counts())


grupos = [dados[dados['Grupo'] == grupo]['AveragScore'] for grupo in dados['Grupo'].unique()]

# Teste ANOVA
f_stat, p_val = stats.f_oneway(*grupos)

print(f"Estatística F: {f_stat}")
print(f"Valor-p: {p_val}")

# Resultado
if p_val < 0.05:
    print("Rejeitamos a hipótese nula: há diferença significativa entre as médias dos grupos.")
else:
    print("Falhamos em rejeitar a hipótese nula: não há diferença significativa entre as médias dos grupos.")
```

```
Grupo
Baixo          42
Médio-baixo    42
Alto           42
Médio-alto     41
Name: count, dtype: int64
Estatística F: 278.3091392607335
Valor-p: 6.895039921255785e-64
Rejeitamos a hipótese nula: há diferença significativa entre as médias dos grupos.
```

Com esses resultados, é possível concluir que existe uma diferença significativa entre as médias dos grupos, baseada na métrica agregada de SafetySecurity, PersonelFreedom e LivingConditions. O valor-p extremamente baixo (6.89e-64) indica que a diferença entre as médias dos grupos é estatisticamente significativa.

5 TREINAMENTO DO MODELO

A regressão linear é uma técnica estatística amplamente utilizada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. No contexto deste projeto, a regressão linear foi aplicada para entender como diferentes fatores influenciam a variável alvo, **AveragScore**. A técnica permite prever o valor da variável dependente com base nas variáveis independentes.

O primeiro passo na implementação da regressão linear foi a preparação dos dados, que envolveu a normalização e codificação das variáveis, garantindo que todas estivessem adequadas para análise. Em seguida, o modelo de regressão linear foi treinado utilizando um subconjunto dos dados, e suas previsões foram compa-

radas aos valores reais da variável alvo. Abaixo, apresentamos o código utilizado para a implementação da regressão linear, após a divisão dos dados, e os resultados obtidos.

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Criar o modelo de Regressão Linear
modelo_regressao = LinearRegression()
# Treinar o modelo com os dados de treino
modelo_regressao.fit(X_train, y_train)

# Fazer previsões com os dados de teste
y_pred = modelo_regressao.predict(X_test)

# Avaliar o modelo
mse = mean_squared_error(y_test, y_pred)  # Erro quadrático médio
r2 = r2_score(y_test, y_pred)  # R-quadrado (coeficiente de determinação)

# Resultados
print(f"Erro Quadrático Médio (MSE): {mse:.2f}")
print(f"Coeficiente de Determinação (R²): {r2:.2f}")

# Coeficientes da Regressão
print("Coeficientes da Regressão Linear:", modelo_regressao.coef_)
print("Intercepto da Regressão Linear:", modelo_regressao.intercept_)
```

5.1 RESULTADOS

```
Erro Quadrático Médio (MSE): 0.00
Coeficiente de Determinação (R²): 1.00
Coeficientes da Regressão Linear: [0.08329498 0.08332064 0.08339341 0.08331529 0.08336579 0.08321716
 0.08328217 0.08340964 0.08332394 0.08338372 0.0833493  0.08330621]
Intercepto da Regressão Linear: 0.002268230502281199
```

- Erro Quadrático Médio (MSE): 0.00

Um MSE de 0 indica que o modelo fez previsões perfeitas, ou seja, não houve erro entre os valores previstos e os valores reais.

- Coeficiente de Determinação (R^2): 1.00

R^2 de 1 indica que o modelo explica 100% da variação dos dados, ou seja, o ajuste do modelo aos dados foi perfeito.

- Coeficientes da Regressão Linear

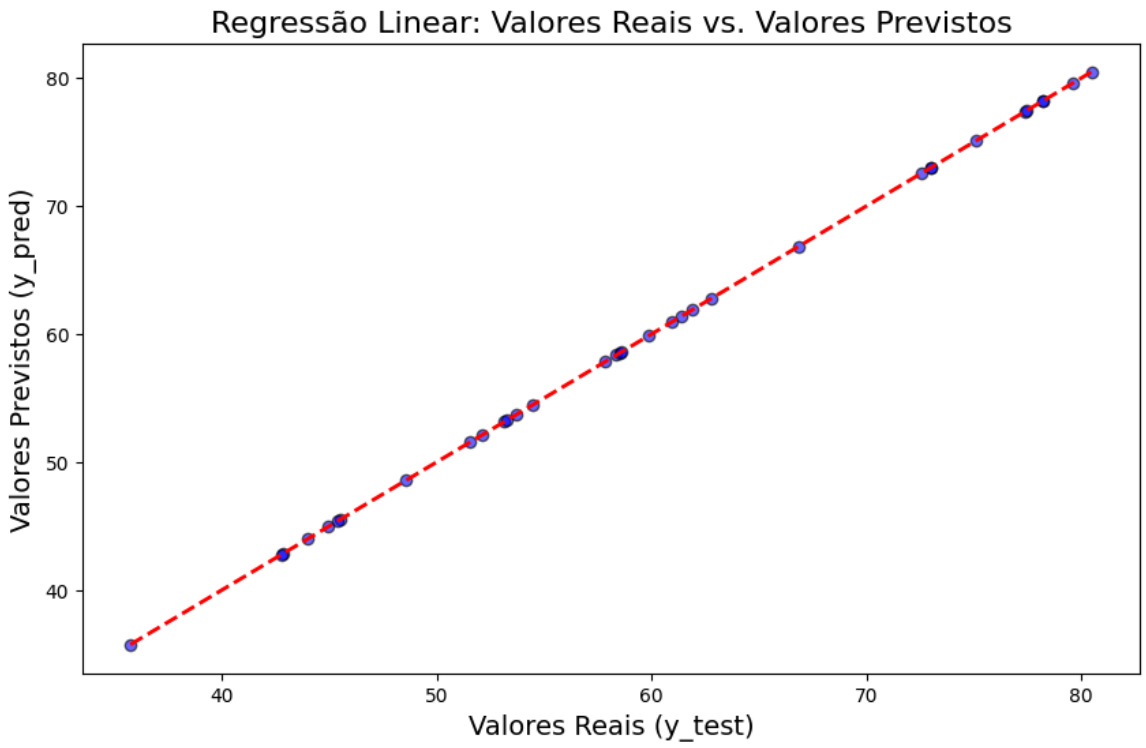
Estes são os pesos que o modelo atribui a cada uma das variáveis independentes do dataset. Cada coeficiente reflete o impacto unitário de sua respectiva variável sobre a variável dependente (AveragScore). Como os coeficientes estão muito próximos, isso sugere que cada variável tem uma contribuição semelhante para a previsão.

- Intercepto da Regressão Linear

O intercepto indica o valor da variável dependente (AveragScore) quando todas as variáveis independentes têm valor zero. Neste caso, o intercepto é pequeno, aproximadamente 0.002, o que indica que, na ausência das variáveis explicativas, a pontuação média seria próxima de zero.

A seguir, um gráfico de dispersão que mostra a relação entre os valores reais e os valores previstos pelo modelo. A linha vermelha pontilhada indica a linha perfeita onde os valores previstos seriam exatamente iguais aos valores reais. Quanto mais perto os pontos estiverem dessa linha, melhor o desempenho do modelo.

Gráfico 4 - Resultado da regressão linear com gráfico de dispersão



5.2 MÉTRICAS UTILIZADAS

Neste projeto, as métricas utilizadas para avaliar o desempenho do modelo de Regressão Linear são o **Erro Quadrático Médio (MSE)** e o **Coeficiente de Determinação (R^2)**. Essas métricas são amplamente utilizadas em problemas de regressão e fornecem uma visão clara sobre a qualidade das previsões do modelo. A seguir, apresentamos uma descrição detalhada dessas métricas, com ênfase no seu papel na análise e no ajuste de modelos de regressão.

1. Erro Quadrático Médio (MSE)

O Erro Quadrático Médio é uma métrica que mede o erro médio ao quadrado entre os valores reais e os valores previstos pelo modelo. O MSE avalia a média das diferenças ao quadrado entre os valores reais e os valores previstos. Ele é sempre positivo e, quanto menor o MSE, melhor o modelo ajustou os dados.

Neste projeto, o MSE é adequado porque permite medir o quão distante as previsões estão dos valores reais. Como o objetivo da regressão linear é fazer previsões precisas, minimizar o MSE é essencial para garantir a qualidade do modelo.

2. Coeficiente de Determinação (R^2)

O Coeficiente de Determinação é uma métrica que indica a proporção da variabilidade total da variável dependente (AveragScore) que é explicada pelas variáveis independentes no modelo. O R^2 varia de 0 a 1, e quanto mais próximo de 1, melhor o modelo está explicando a variabilidade dos dados. Um R^2 de 0 indica que o modelo não explica nenhuma variabilidade, enquanto um R^2 de 1 indica que o modelo explica toda a variabilidade dos dados.

O Coeficiente de Determinação é uma métrica importante porque ajuda a entender o quanto o modelo está capturando da variabilidade total da variável dependente. No contexto da regressão linear, um alto valor de R^2 indica que o modelo está fazendo boas previsões.

6 MODELO DE NEGÓCIOS

O modelo de negócios busca transformar a análise e os insights de prosperidade em consultoria estratégica e produtos de dados. A ideia central é oferecer soluções analíticas e estratégicas para governos, ONGs e empresas de impacto social interessadas em aumentar os índices de prosperidade em áreas críticas.

Proposta:

- Consultoria em Prosperidade: Serviços personalizados de consultoria para apoiar políticas públicas e programas sociais focados nos indicadores do índice de prosperidade.
- Relatórios Customizados: Relatórios personalizados que destacam áreas específicas de prosperidade, com recomendações de ação, direcionados às necessidades de cada cliente.

Segmentos de Clientes:

- Governos e agências governamentais interessadas em aumentar a prosperidade em áreas específicas.
- ONGs e instituições de desenvolvimento que precisam de dados confiáveis para estruturar projetos de impacto.
- Corporações que buscam investir em responsabilidade social e precisam de uma base analítica para seus programas.

Indicadores de Sucesso:

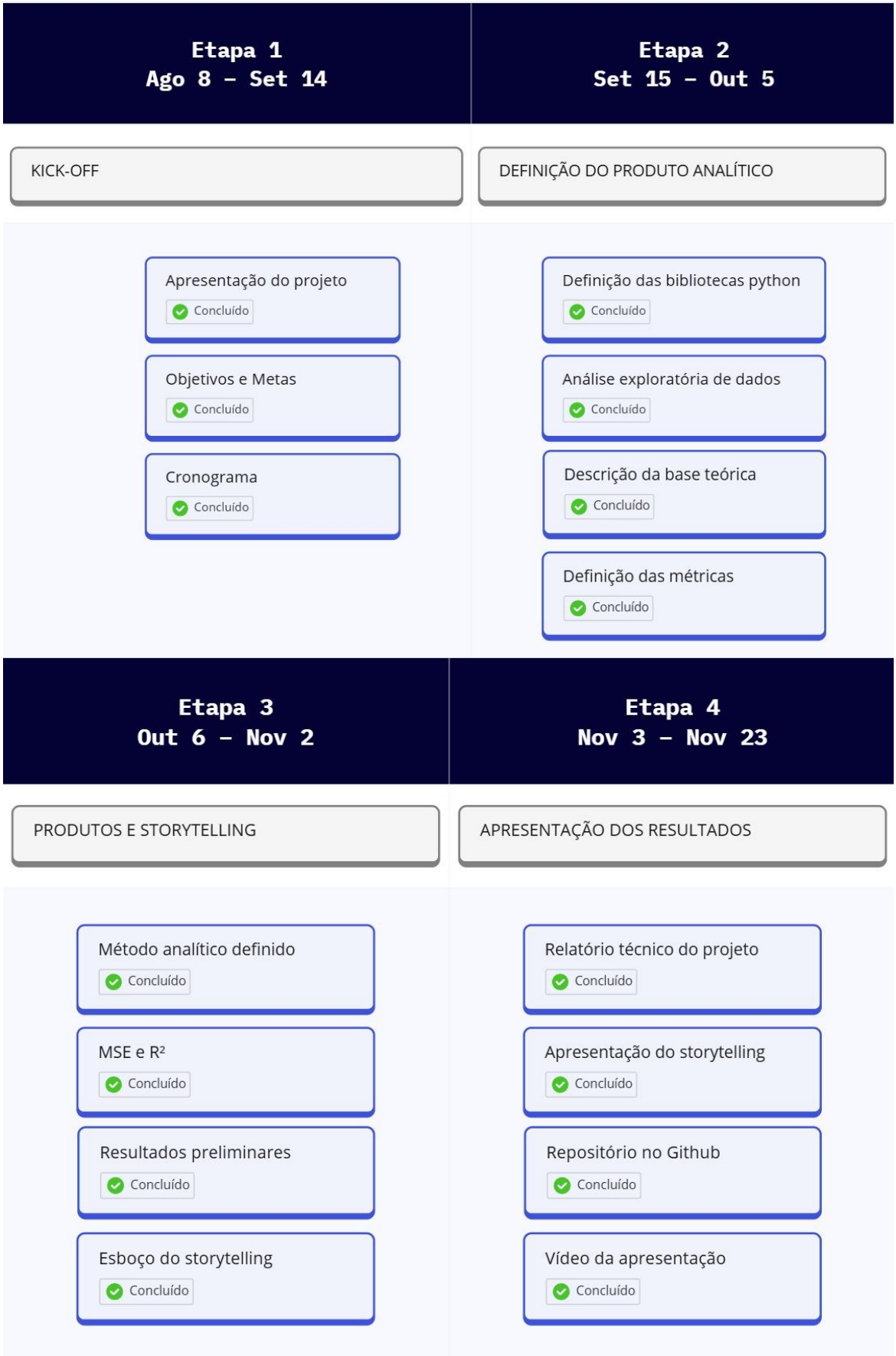
- Crescimento no número de contratos de consultoria e relatórios.
- Impacto positivo dos programas nas áreas de atuação dos clientes, medido pela evolução nos índices de prosperidade.

7 GITHUB

Acesso ao repositório do projeto:

https://github.com/GrupoProjetoAplicado/Projeto-Aplicado-II

CRONOGRAMA


Etapa 1
Ago 8 – Set 14
Etapa 2
Set 15 – Out 5
KICK-OFF
DEFINIÇÃO DO PRODUTO ANALÍTICO
Apresentação do projeto
Concluído
Objetivos e Metas
Concluído
Cronograma
Concluído
Definição das bibliotecas python
Concluído
Análise exploratória de dados
Concluído
Descrição da base teórica
Concluído
Definição das métricas
Concluído
Etapa 3
Out 6 – Nov 2
Etapa 4
Nov 3 – Nov 23
PRODUTOS E STORYTELLING
APRESENTAÇÃO DOS RESULTADOS
Método analítico definido
Concluído
MSE e R²
Concluído
Resultados preliminares
Concluído
Esboço do storytelling
Concluído
Relatório técnico do projeto
Concluído
Apresentação do storytelling
Concluído
Repositório no Github
Concluído
Vídeo da apresentação
Concluído

MEMBROS

Giovanna Sobral da Silva

Carla Pollastrini

Felipe Akira Fukue