

Relatório Sobre Análise de Dados do Desastre do Titanic

Uma imagem com transporte, embarcação, Arquitetura naval, navio

Descrição gerada automaticamente

Atec - TPSICAS1123 - 5417

Relatório realizado por: João Pedro Fernandes Silva, Ricardo Lopes da Conceição e Diogo Baptista Louro

Formador: Nelson Alexandre Santos

Índice

[Introdução 3](#_Toc184317348)

[Estrutura dos dados 3](#_Toc184317349)

[Leitura e Exploração dos Dados 4](#_Toc184317350)

[Limpeza e Pré-Processamento 5](#_Toc184317351)

[Visualização de dados 10](#_Toc184317352)

[Análise Adicional – Relação entre Família e Sobrevivência 16](#_Toc184317353)

[Análise Adicional – Comparação por Porto de Embarque 17](#_Toc184317354)

[Exportação dos resultados 18](#_Toc184317355)

[Utilização do Git 19](#_Toc184317356)

[Conclusão 20](#_Toc184317357)

Índice de Figuras

[Figura 1 - Utilização de .head, .tail e .describe 4](#_Toc184317358)

[Figura 2 - Resumo estatístico 4](#_Toc184317359)

[Figura 3 - Info geral sobre o DataFrame 4](#_Toc184317360)

[Figura 4 - Valores nulos existentes, por coluna 5](#_Toc184317361)

[Figura 5 - Valores nulos após tratamento 5](#_Toc184317362)

[Figura 6 - Tipo de dados 6](#_Toc184317363)

[Figura 7 - Idades em milissegundos desde o Epoch 6](#_Toc184317364)

[Figura 8 - Taxa de sobrevivência por classe e género 7](#_Toc184317365)

[Figura 9 - Media de idade dos sobreviventes 7](#_Toc184317366)

[Figura 10 - Idade média por classe e sobrevivência 8](#_Toc184317367)

[Figura 11 - Distribuição da Idade por Sobrevivência 8](#_Toc184317368)

[Figura 12 - Gráfico de Linha: Representa a distribuição dos sobreviventes por classe e sexo 10](#_Toc184317369)

[Figura 13 - Gráfico de Dispersão: Representa a correlação entre as variáveis Age, Fare e Survived 11](#_Toc184317370)

[Figura 14 - Gráfico de Dispersão: Representa a correlação entre Idade e Sobrevivência 11](#_Toc184317371)

[Figura 15 - Gráfico de Dispersão: Representa a correlação entre a Tarifa vs Sobrevivência 12](#_Toc184317372)

[Figura 16 - Gráfico de Barras: Distribuição das Idades 13](#_Toc184317373)

[Figura 17 - Gráfico de Barras: Distribuição de Tarifas 14](#_Toc184317374)

[Figura 18 - Gráfico de Barras: Distribuição da Sobrevivência 15](#_Toc184317375)

[Figura 19 - Prova de criação da nova coluna Tamanho\_Familia 16](#_Toc184317376)

[Figura 20 - Gráfico de Barras: Relação entre o Tamanho das Familias e a Sobrevivência 16](#_Toc184317377)

[Figura 21 - Coluna Embarked 17](#_Toc184317378)

[Figura 22 - Gráfico de Barras: Taxa de sobrevivência por Porto de Embarque 17](#_Toc184317379)

[Figura 23 - Ficheiro .csv com dados tratados 18](#_Toc184317380)

[Figura 24 - Ficheiro final .csv criado 18](#_Toc184317381)

[Figura 25 - Amostra dos Ficheiros.CSV 18](#_Toc184317382)

[Figura 26 - Utilização do Git 19](#_Toc184317383)

Introdução

Neste relatório analisámos um conjunto de dados utilizando Python e bibliotecas como Pandas para manipulação e análise de dados e Matplotlib para visualizações gráficas. O conjunto de dados fornecido contém informações sobre os passageiros como a idade, sexo, classe de embarque e se sobreviveram ou não.

O objetivo deste projeto é realizar uma análise com uma exploração inicial dos dados e apresentação de insights relevantes. O relatório inclui etapas de limpeza e pré-processamento dos dados, criação de métricas adicionais e visualizações gráficas.

Estrutura dos dados

Os dados fornecidos contêm 12 colunas sobre as características dos passageiros e a sobrevivência dos mesmos. Analisando os dados temos o seguinte:

* PassengerId: identificador único;
* Survived: 1 (sobreviveu) ou 0 (não sobreviveu);
* Pclass: classe do passageiro (1ª, 2ª ou 3ª classe);
* Name: nome do passgeiro;
* Sex: género do passageiro;
* Age: idade dos passageiros;
* SibSp: número de irmãos e esposo/a do passageiro;
* Parch: números de pais e filhos do passageiro;
* Ticket: número do bilhete do passageiro;
* Fare: taxa cobrada por passageiro;
* Cabin: número da cabine do passageiro;
* Embarked: porto de embarque C, Q ou S.

O dataset foi carregado com pd.read\_csv. Os métodos .head(), .tail(), .describe() e .info() foram usados para analizar a estrutura de dados, identificar valores nulos e compreender o formato das colunas.

Leitura e Exploração dos Dados

Os dados foram carregados com pd.read\_csv() e inspecionados utilizando .head, .tail, .describe e .info respetivamente, obtendo os seguintes resultados:

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Utilização de .head, .tail e .describe

Uma imagem com texto, captura de ecrã

Descrição gerada automaticamente

Figura - Resumo estatístico

Uma imagem com texto, captura de ecrã, menu, Tipo de letra

Descrição gerada automaticamente

Figura - Info geral sobre o DataFrame

Limpeza e Pré-Processamento

Para tratar dos valores nulos e ajustar os dados na coluna Age os valores ausentes foram preenchidos com a mediana das idades, a coluna Cabin foi descartada na análise inicial devido à alta quantidade de valores ausentes e na coluna Embarked os valores ausentes foram preenchidos com o valor mais frequente.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Valores nulos existentes, por coluna

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Valores nulos após tratamento

Relativamente ao tipo de dados, consideramos não haver a necessidade de converter qualquer tipo.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Tipo de dados

Nota: PassengerId, Survived, Pclass: Está como int64, o que está correto. Como é um identificador único, não precisa de conversão. Age, Flare: Estão corretamente como float64. Name, Ticket, Cabin, Embarked, Sex: Mantêm o tipo object (string).

Depois criamos uma coluna chamada *Idade\_Milissegundos*, onde as idades foram convertidas para milissegundos.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Idades em milissegundos desde o Epoch

Nota: Para idades inferiores à data do Epoch (1970-01-01) decidimos colocar o número inteiro pois apresentava número negativo em milissegundos.

Análise e Manipulação dos Dados

Realizamos algumas análises com as funções do Pandas, como *groupby* para calcular a taxa de sobrevivência por classe e género. Foi possível observar que a classe não interferiu na taxa de sobrevivência. Relativamente ao género, as mulheres apresentaram uma taxa de sobrevivência absoluta com nenhum homem a sobreviver.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

Figura - Taxa de sobrevivência por classe e género

De seguida analisámos qual a idade média de um sobrevivente o que nos fez concluir que a probabilidade de sobreviver é mais favorável para pessoas mais jovens.

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura - Media de idade dos sobreviventes

Na próxima analise, vamos interpretar os dados da classe vs sobrevivência e perceber se o status e/ou a idade prevaleceu sobre a decisão dos que sobreviviam.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente

Figura - Idade média por classe e sobrevivência

Uma imagem com diagrama, captura de ecrã, Gráfico, texto

Descrição gerada automaticamente

Figura - Distribuição da Idade por Sobrevivência

Interpretação dos dados:

Classe 1 (Pclass = 1):

* Não sobreviveram (Survived = 0): A idade média dos passageiros que não sobreviveram nesta classe foi de aproximadamente 38,86 anos.
* Sobreviveram (Survived = 1): A idade média dos passageiros que sobreviveram nesta classe foi de aproximadamente 40,76 anos.

Conclusão: Pode indicar que passageiros mais velhos tinham maior probabilidade de sobrevivência nessa classe, ou que havia mais passageiros mais velhos sobreviventes na 1ª classe.

Classe 2 (Pclass = 2):

* Não sobreviveram (Survived = 0): A idade média foi de 30,69 anos.
* Sobreviveram (Survived = 1): A idade média foi de 24,46 anos.

Conclusão: Pode indicar que passageiros mais jovens tiveram maior probabilidade de sobreviver nesta classe.

Classe 3 (Pclass = 3):

* Não sobreviveram (Survived = 0): A idade média foi de 25,37 anos.
* Sobreviveram (Survived = 1): A idade média foi de 24,27 anos.

Conclusão: Os sobreviventes tendem a ser ligeiramente mais jovens.

Portanto, estes padrões podem ser influenciados por fatores como:

* Acesso a botes salva-vidas (maior nas classes mais altas).
* Prioridades dadas durante o resgate (potencialmente favorecendo crianças e mulheres).

# Visualização de dados

No gráfico seguinte podemos ter uma visão mais clara sobre a classe e sexo dos sobreviventes onde imediatamente podemos concluir que relativamente às mulheres houve um número muito baixos de sobreviventes de segunda classe e um número mais alto para mulheres de terceira classe.

Uma imagem com file, Gráfico, diagrama, texto

Descrição gerada automaticamente

Figura - Gráfico de Linha: Representa a distribuição dos sobreviventes por classe e sexo

Analisando as idades com a tarifa e sobrevivência podemos observar que a tarifa não teve impacto na taxa de sobrevivência pois maior parte dos sobreviventes pagaram uma tarifa Fare menor que 100. O número 0 representa os não-sobreviventes enquanto que o número 1 representa os sobreviventes.

Uma imagem com captura de ecrã, file, diagrama, texto

Descrição gerada automaticamente

Figura - Gráfico de Dispersão: Representa a correlação entre as variáveis Age, Fare e Survived

Uma imagem com texto, captura de ecrã, file, Gráfico

Descrição gerada automaticamente

Figura - Gráfico de Dispersão: Representa a correlação entre Idade e Sobrevivência

Uma imagem com texto, file, captura de ecrã, Gráfico

Descrição gerada automaticamente

Figura - Gráfico de Dispersão: Representa a correlação entre a Tarifa vs Sobrevivência

No histograma seguinte podemos visualizar melhor a distribuição de idades sobre a taxa de sobrevivência. O número 0 representa os não-sobreviventes enquanto que o número 1 representa os sobreviventes.

Uma imagem com file, Gráfico, diagrama, captura de ecrã

Descrição gerada automaticamente

Figura - Gráfico de Barras: Distribuição das Idades

Analise: Pode-se concluir que o grosso das pessoas com idade entre 20-30 anos foi o que não sobreviveu, e ainda com 40 anos. Por outro lado, foi entre os 0-30 anos que mais sobreviventes se pode observar, e ainda na casa dos 45-50 anos.

Neste histograma podemos visualizar a distribuição das tarifas sobre a taxa de sobreviventes.

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Descrição gerada automaticamente

Figura - Gráfico de Barras: Distribuição de Tarifas

Analise: O número de não sobreviventes que adquiriram uma tarifa mais barata é bastante visível, enquanto quem pagou uma tarifa mais cara, teve mais chances de sobrevier.

E neste gráfico podemos observar melhor a diferença do número de passageiros que sobreviveram e o número de passageiros que não sobreviveram em que claramente houve muitos mais passageiros que não sobreviveram.

Uma imagem com texto, captura de ecrã, diagrama, Gráfico

Descrição gerada automaticamente

Figura - Gráfico de Barras: Distribuição da Sobrevivência

# Análise Adicional – Relação entre Família e Sobrevivência

Uma análise adicional que fizemos, a qual necessitou de se criar uma nova coluna, foi explorar qual a influência do tamanho da família na taxa de sobrevivência e obtemos o seguinte gráfico que mostra que um número de membros familiares de 4 e 6 apresentam uma taxa de sobrevivência relativamente mais elevada que passageiros com 1, 5 e 7 membros familiares com os restantes valores por volta de 50% de probabilidade de sobrevivência.

Uma imagem com texto, captura de ecrã, software, Tipo de letra

Descrição gerada automaticamente

Figura - Prova de criação da nova coluna Tamanho\_Familia

Uma imagem com texto, captura de ecrã, diagrama, file

Descrição gerada automaticamente

Figura - Gráfico de Barras: Relação entre o Tamanho das Familias e a Sobrevivência

# Análise Adicional – Comparação por Porto de Embarque

Ao analisarmos o gráfico obtido da taxa de sobrevivência por porto de embarque concluímos que os passageiros que embarcaram em Queenstown tiveram uma maior probabilidade de sobreviver seguido do porto de Cherbourg e por fim os passageiros de Southampton tiveram a menor probabilidade de sobrevivência.

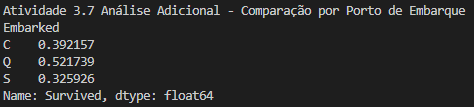


Figura - Coluna Embarked

Uma imagem com texto, captura de ecrã, diagrama, file

Descrição gerada automaticamente

Figura - Gráfico de Barras: Taxa de sobrevivência por Porto de Embarque

# Exportação dos resultados

Ao longo do projecto, foi necessário exportar para ficheiro Excel, aquando da conclusão das atividades que resultou na seguinte criação de ficheiros .csv:

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura - Ficheiro .csv com dados tratados



Figura - Ficheiro final .csv criado

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Figura - Amostra dos Ficheiros.CSV

# Utilização do Git

Neste trabalho foi utilizada a plataforma Git, para que todos os elementos do grupo pudessem trabalhar em simultâneo, e atualizar o projeto conforme as funcionalidades criadas:

Uma imagem com texto, captura de ecrã, software, Software de multimédia

Descrição gerada automaticamente

Figura - Utilização do Git

# Conclusão

Neste relatório analisámos os dados sobre os passageiros do Titanic e desta forma conseguimos identificar padrões que influenciaram a sobrevivência ao naufrágio de 1912.

Nos dados analisados encontramos claras evidências da política "mulheres e crianças primeiro” com taxas de sobrevivência mais altas para mulheres e passageiros mais jovens. Os passageiros da 3ª e 1ª classe tiveram mais sobreviventes que 2ª classe, onde destacamos a posição e acessibilidade aos botes salva-vidas.

Relativamente a tarifas não houve um impacto o sendo que os que tiveram tarifas inferiores a 100 tinham maior taxa de sobrevivência.

Com esta análise, foi possível perceber que os fatores sociais e económicos tiveram a sua influência na sobrevivência deste desastre.

Por fim, achamos que este trabalho foi uma boa experiência, que nos permitiu consolidar os conceitos da Programação em Python, e como é uma mais valia na Analise de Dados.