UNIVERSIDADE FEDERAL DO PARANÁ

ERICKSON LEON KOVALSKI

PROJETO FINAL: SPACESHIP TITANIC DATASET

Relatório de projeto apresentado à disciplina de Aprendizado de Máquinas como requisito de avaliação na disciplina *Aprendizado de Máquina* CE329

Professor Eduardo Vargas Ferreira

CURITIBA
2023

Introdução

A análise de dados e a aplicação de modelos preditivos desempenham um papel fundamental nas ciências da informação, proporcionando ferramentas que orientam decisões em diversas áreas. Uma competição que exemplifica a integração dessas práticas é a Kaggle Spaceship Titanic, que, por meio de técnicas avançadas de aprendizado de máquina, busca antecipar eventos em uma situação hipotética envolvendo a Spaceship Titanic no ano 2912.

Para contextualizar essa abordagem preditiva, é imperativo recordar o desastre do Titanic em 1912. O naufrágio do transatlântico Titanic resultou em perdas significativas de vidas humanas devido à colisão com um iceberg. Este evento traumático motivou uma análise extensa e contínua sobre os fatores que influenciaram a sobrevivência dos passageiros. A análise desses dados históricos do Titanic frequentemente emprega técnicas de Estatística e Aprendizado de Máquina para entender os padrões e correlações entre diferentes variáveis e o resultado, ou seja, se um passageiro sobreviveu ou não. Regressão e classificação são duas abordagens fundamentais na predição algorítmica. A regressão lida com a previsão de valores contínuos, enquanto a classificação aborda a categorização de dados em classes distintas. A variável resposta em tal contexto é binária – a sobrevivência (ou não) de um passageiro. No contexto da competição Kaggle Spaceship Titanic, trata-se de um problema de classificação binária, uma vez que se busca prever se um passageiro foi ou não transportado para uma dimensão alternativa após o incidente. Neste cenário, a competição assume um papel significativo não apenas como uma extensão da análise do Titanic, mas como uma plataforma inovadora para testar e aprimorar as habilidades dos praticantes de ciência de dados.

Contextualização do Problema

A Spaceship Titanic, uma nave interplanetária de passageiros, colidiu com uma anomalia espaço-temporal, resultando no transporte de quase metade dos passageiros para uma dimensão alternativa. O desafio proposto pela competição Kaggle Spaceship Titanic é formular modelos preditivos capazes de identificar quais passageiros foram transportados para essa dimensão alternativa com base em registros recuperados da nave danificada. A natureza desse problema é

classificatória binária, onde o objetivo é determinar se um passageiro específico foi ou não transportado.

Características do Dataset

O conjunto de dados fornecido compreende dois arquivos principais: o conjunto de treino (train.csv) e o conjunto de teste (test.csv). Ambos contêm informações pessoais dos passageiros, bem como variáveis que podem influenciar a probabilidade de serem transportados para a dimensão alternativa. Essas variáveis englobam, entre outros aspectos, o planeta de origem, a decisão de entrar em sono criogênico, o número da cabine, o destino pretendido, a idade, serviços VIP utilizados e os gastos em amenidades da nave. Destaca-se que uma quantidade considerável de informações está ausente e requer tratamento para possibilitar avanços na análise.

Técnicas de Análise de Dados

A análise desses dados requer diversas etapas, inicialmente compreende uma Análise Exploratória de Dados, onde se realiza uma inspeção detalhada das distribuições das variáveis, identificação de padrões e relações entre as características. Além disso, estratégias de pré-processamento são aplicadas para lidar com valores ausentes e outliers, garantindo a integridade e qualidade dos dados. Em seguida, a etapa de Engenharia de Características se destaca, visando criar novos atributos derivados das informações existentes que podem fornecer mais poder preditivo ao modelo. Por fim, a construção do modelo preditivo baseia-se no uso de algoritmos de aprendizado de máquina mais comuns, sendo o Random Forest Classifier e Logarithmic Regression, exemplos explorado neste contexto. O ajuste de hiperparâmetros é realizado para otimizar a performance do modelo, buscando atingir uma acurácia máxima na predição dos eventos de transporte para a dimensão alternativa.

Materias e Métodos

Descrição do Conjunto de Dados:

Os dados empregados neste estudo foram adquiridos por meio de dois principais arquivos disponíveis no Kaggle, uma plataforma reconhecida por competições e compartilhamento de conjuntos de dados voltados para ciência de dados e aprendizado de máquina. Esses arquivos, intitulados "train.csv" e "test.csv", constituem conjuntos sintéticos, gerados especificamente para a competição em questão. Vale ressaltar que tais dados não resultam de observações diretas de fenômenos do mundo real; entretanto, apresentam semelhanças com o manifesto de passageiros do transatlântico Titanic.

Variáveis nos Conjuntos de Treino e Teste:

O conjunto de treino, denominado "train.csv", contém informações pessoais dos passageiros e variáveis que são potenciais influenciadores na probabilidade de serem transportados para a dimensão alternativa. As variáveis incluem, mas não se limitam a:

- *PassengerId*: Identificação única para cada passageiro. Tem a característica de possuir a identificação da família e do indivíduo no formato XXXX_YY.
- *HomePlanet*: Planeta de origem do passageiro. São três valores possíveis.
- *CryoSleep*: Indicação se o passageiro optou por entrar em sono criogênico, variável binária.
- *Cabin*: Número da cabine do passageiro. Possui número, lado e corredor codificado neste campo.
- *Destination*: Destino pretendido do passageiro. São três valores possíveis.
- *Age*: Idade do passageiro.
- *VIP*: Indicação se o passageiro utilizou serviços VIP, variável binária.
- *RoomService*, *FoodCourt*, *ShoppingMall*, *Spa*, *VRDeck*: Gastos em diferentes amenidades da nave, são variáveis numéricas.

- *Name*: Nome completo do passageiro.
- *Transported*: Variável alvo, indicando se o passageiro foi transportado para outra dimensão.

O conjunto de teste, denominado "test.csv", possui uma estrutura semelhante ao conjunto de treino, excluindo a variável alvo "Transported", que deve ser prevista com base no modelo desenvolvido com o conjunto de treino.

Estudo Exploratório e Descritivo

Aproximadamente metade dos passageiros a bordo as espaçonave foram vítimas do acidente que os transportou a outra dimensão (Gráfico 1), o conjunto de dados, portanto, parece ser bem balanceado, o que facilita o trabalho a ser desenvolvido, exigindo técnicas estatísticas menos complexas para esta variável.

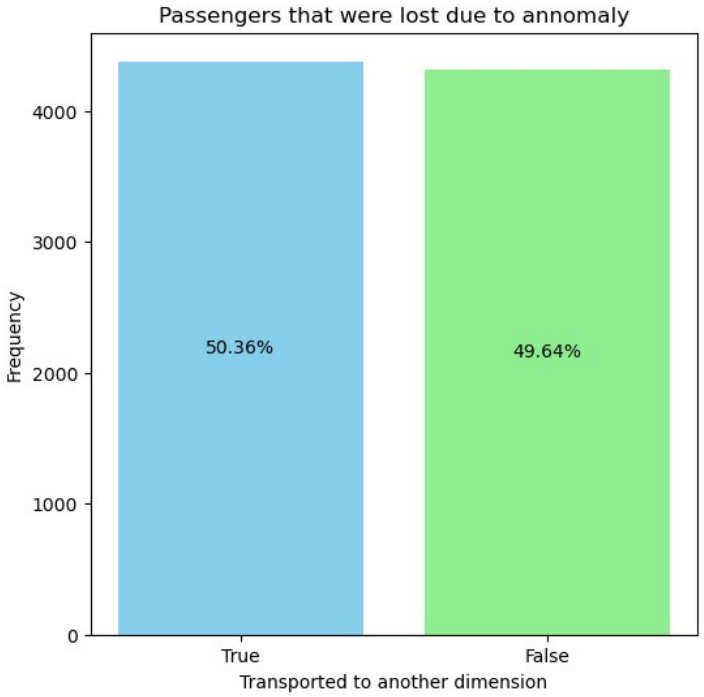

Gráfico 1: Passageiros Transportados a outra dimensão

Quanto à origem dos passageiros (Gráfico 2) a maior parte, 52,94%, parece ter vindo da Terra, enquanto 24,51% vêm da lua Europa enquanto que a menor parcela, de 20,23% dos passageiros é originária das colônias de Marte. Isso não inclui uma parcela de passageiros cuja origem está faltando na base de dados.

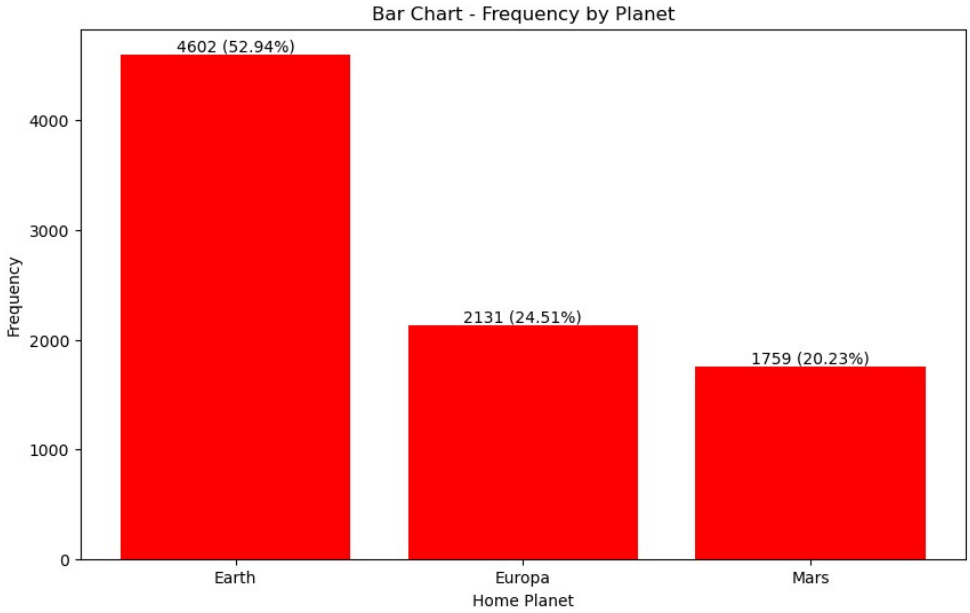

Gráfico 2: Origem dos Passageiros

Em contrapartida, o destino mais comum é TRAPPIST-1e, com 68% dos passageiros para os quais temos esses dados escolhendo esse destino. O segundo mais comum é 55 Cancri-e, escolhido por 20,7% dos passageiros, e o menos comum foi PSO j318.5-22, com apenas pouco mais de 9% (Gráfico 3). A porcentagem restante para completar o total se deve a dados ausentes.

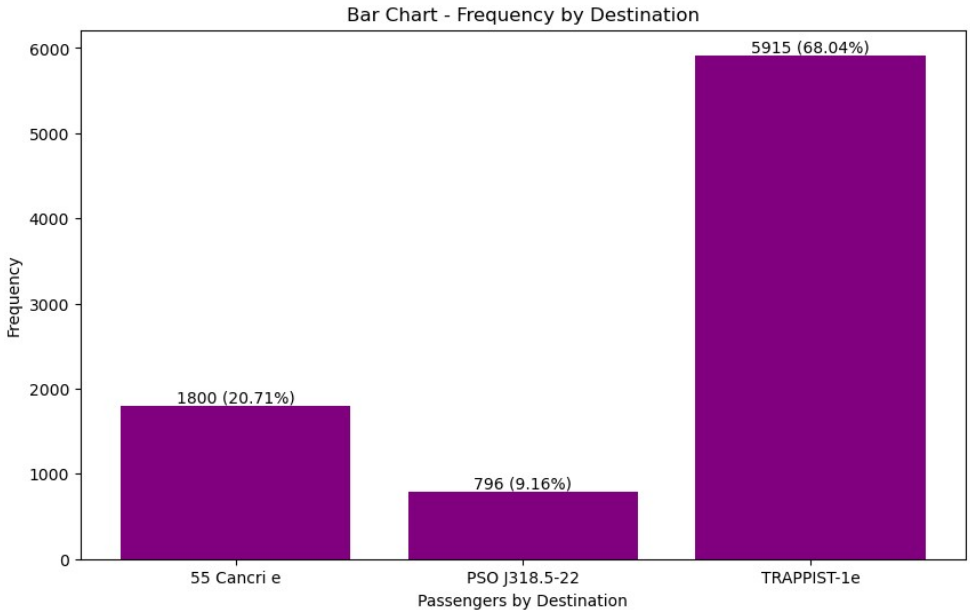

Gráfico 3: Destino da viagem

O método preferido de viagem na espaçonave é por meio do sono criogênico, utilizado por 5439 passageiros em nossa lista (Gráfico 4), sem contar dados ausentes. Isso representa aproximadamente 62,5% da lista de passageiros. Outros

3037 passageiros (cerca de 35%) preferiram permanecer acordados durante o tempo de viagem, aproveitando as instalações disponíveis na espaçonave para passar o tempo. Dados ausentes afetam 2,5% dos dados nesta categoria.

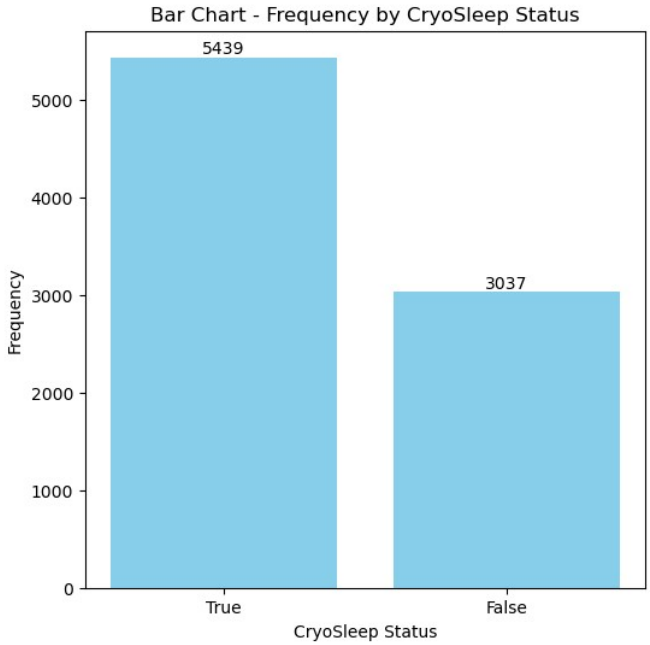

Gráfico 4: Passageiros em sono criogênico

A idade média do passageiro é de 28,8 anos, e a idade mediana é de 27 anos. Um grande número de passageiros tem a idade marcada como 0 e 1, o que pode indicar a ausência de dados. Como observado anteriormente, o passageiro mais idoso a bordo tem 79 anos (Gráfico 5).

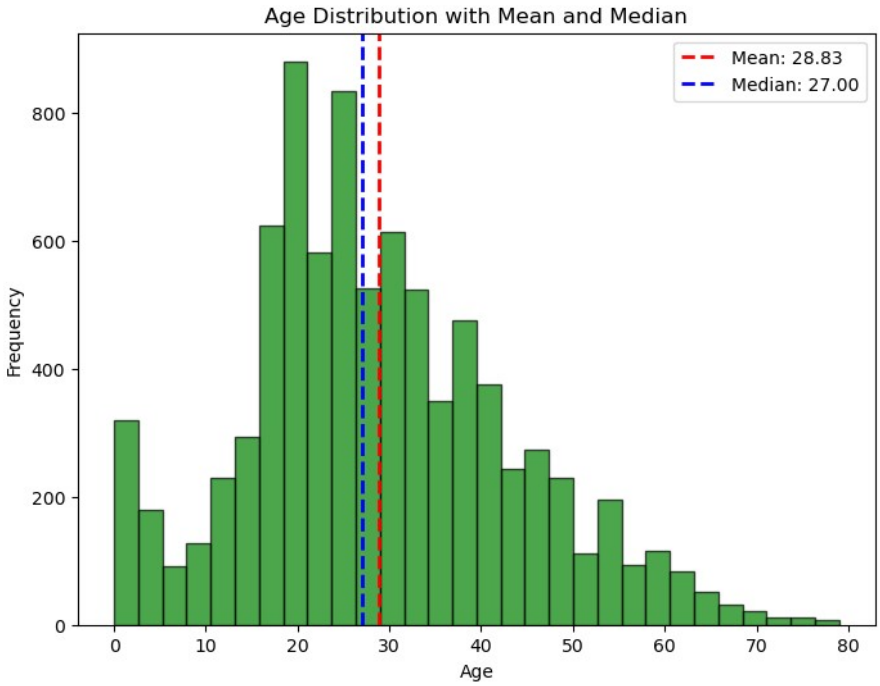

Gráfico 5: Idade dos passageiros segundo o manifesto

Os dados ausentes nesta categoria (Gráfico 6) são ligeiramente inferiores aos de outras categorias, totalizando apenas 2%. No entanto, há um número significativo de passageiros cuja idade está registrada como 0 ou 1. Isso pode contribuir para parte dos dados ausentes, uma vez que a Idade é a categoria com o menor número de valores *NaN*. Nesse cenário, precisamos calcular duas médias para preenchimento dos dados ausentes: uma para adultos e outra para crianças, excluindo esses valores duvidosos. Acreditamos que essa abordagem seja justificada devido ao grande número de idades marcadas como zero ou um. A remoção desses valores pode resultar potencialmente em menor viés em comparação com a manutenção deles como estão.

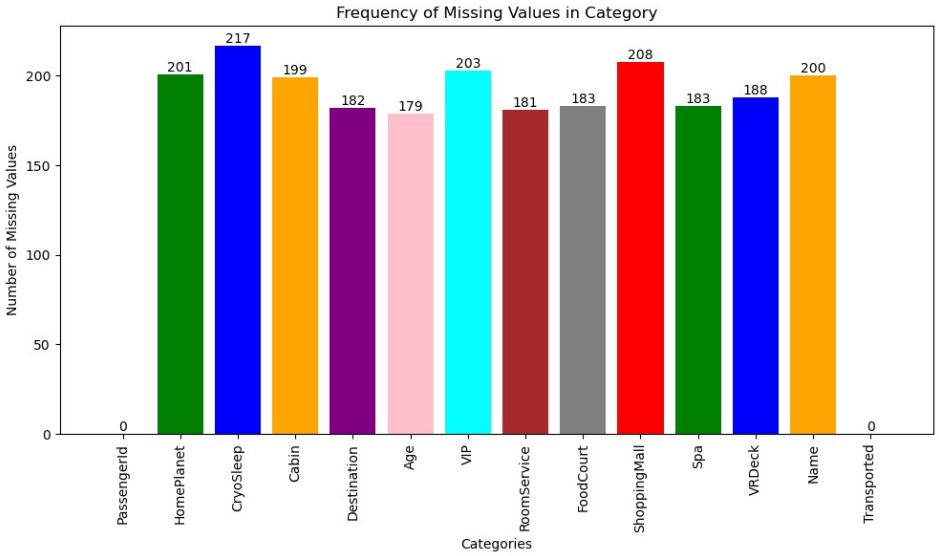

Gráfico 6: Dados faltantes por categoria

No gráfico 7, podemos notar que dos 8693 passageiros, apenas 199 adquiriram o serviço VIP, representando apenas 2,3% dos passageiros. Nesta categoria, identificamos 203 células com dados ausentes.

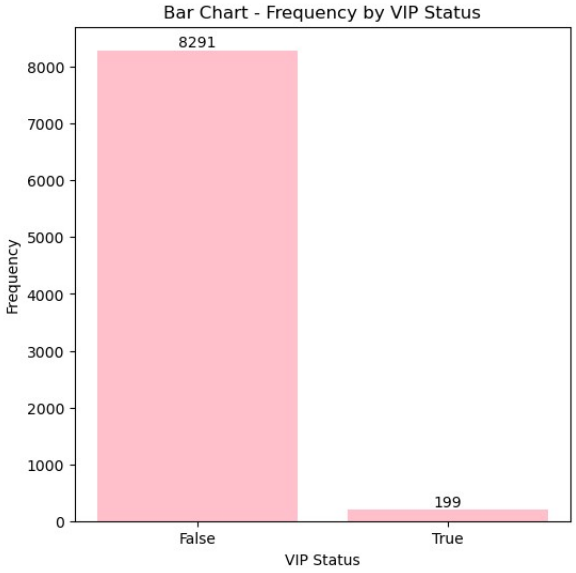

Gráfico 7: Passageiros VIP

Em relação às despesas, a maioria dos passageiros tem gastos muito baixos a bordo da nave espacial, mas alguns são grandes gastadores (*outliers*). Passageiros com menos de 13 anos não tiveram nenhum gasto: isso faz sentido, pois pode estar sob o cartão-chave de seus pais. Como esperado, os passageiros VIP em geral têm despesas mais distribuídas em amenidades do que a população

em geral. Nenhum dos passageiros em sono criogênico teve despesas em qualquer uma das amenidades (Gráfico 8)

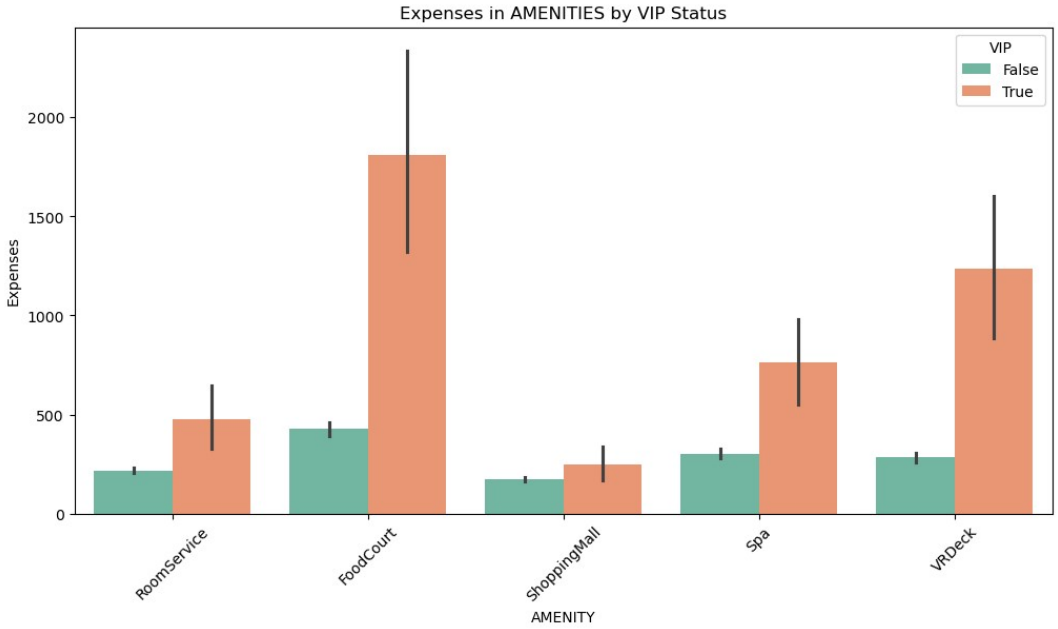

Gráfico 8: Gasto com as amenidades a bordo

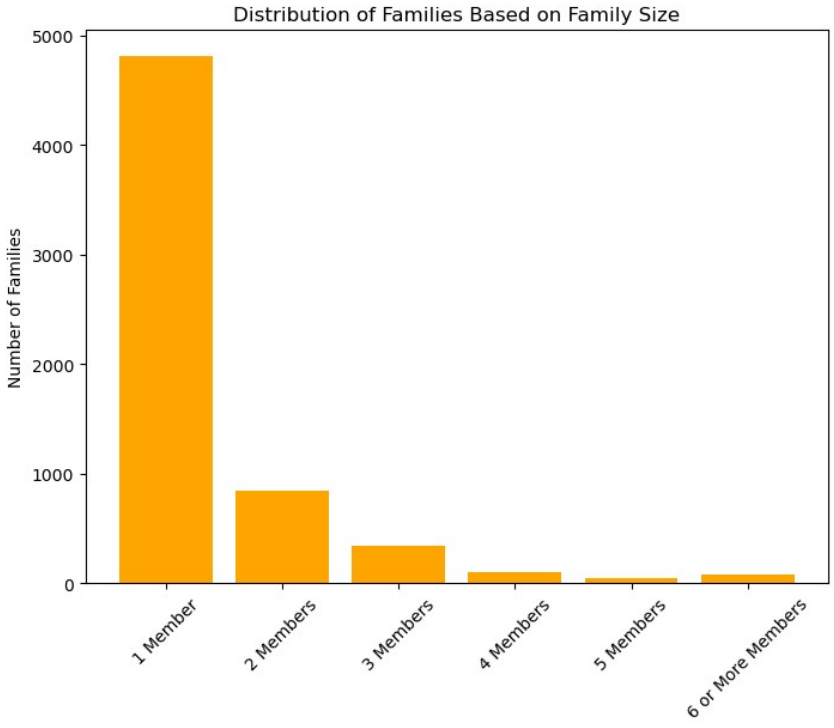

Gráfico 9: Tamanho das Famílias

Os relacionamentos familiares entre os passageiros podem ser determinados a partir da coluna "PassengerId". O código pode ser dividido em duas partes: uma parte é comum ao grupo familiar, e a outra parte identifica o indivíduo dentro da família. Portanto, um número de família único indica que o indivíduo está viajando

sozinho. O Gráfico 9 sugere que a maioria das pessoas estava viajando sozinha, possivelmente se mudando em busca de oportunidades nas colônias.

Pré-processamento de Dados

Para preparar os dados para análise, foram implementadas diversas etapas visando a limpeza e organização adequada das informações:

1.Tratamento de Valores Ausentes: Identificaram-se e trataram-se os valores ausentes presentes nos conjuntos de treino e teste. Nas variáveis categóricas, os valores faltantes foram imputados utilizando a moda, enquanto nas variáveis numéricas, optou-se pela imputação da média para garantir consistência e preservação da distribuição. Foram feitos ainda vários experimentos, que constam ainda no caderno jupyter, sem muito sucesso porém.

2. Tratamento de Outliers: Identificaram-se outliers em variáveis específicas, como "RoomService," "FoodCourt," "ShoppingMall," "Spa," e "VRDeck," relacionadas aos gastos nas amenidades da nave. Utilizou-se uma abordagem de truncamento, limitando os valores extremos a um percentil específico (99%) para preservar a distribuição dos dados.

3. Criação de Novas Variáveis: Com base na análise exploratória, novas variáveis foram criadas para capturar informações relevantes, como a categoria do deck da cabine, posição no deck, total de gastos, média de gastos por deck, número de familiares a bordo, categoria de tamanho da família, entre outras.

4. Remoção de Variáveis Não Relevantes: Variáveis que não contribuíram significativamente para a análise foram removidas, reduzindo a dimensionalidade do conjunto de dados e simplificando o modelo final.

Essas etapas permitiram garantir a consistência e confiabilidade dos dados, preparando para a análise exploratória e a construção de modelos preditivos.

Engenharia de Recursos (Feature Engineering):

Durante o processo de engenharia de recursos, foram criadas novas variáveis a partir das existentes para capturar informações relevantes e potencialmente impactar positivamente os resultados da análise preditiva. Algumas das principais criações incluem:

1. *Criação de Variáveis Relacionadas à Cabine*: Extraída da variável original "Cabin," que representam o corredor e lado da cabine.

Essas variáveis foram introduzidas para explorar padrões relacionados à localização das cabines e sua possível influência na probabilidade de transporte para outra dimensão. A decisão de criar essas variáveis baseou-se na suposição de que a posição no navio poderia ter impacto nos eventos hipotéticos.

2. *Criação de Variáveis Relacionadas à Família*: extração de relação de parentesco a partir do PassengerId possibilitou identificar passageiros que viajavam sozinhos ou em grupos familiares. A criação dessas variáveis visa entender melhor a dinâmica familiar a bordo, considerando tanto o número total de familiares quanto sua distribuição em categorias de tamanho. Essas informações podem ser relevantes para imputar valores faltantes.

3. *Criação de Variáveis Relacionadas a Gastos*: Foram feitos dois experimentos, um relacionando em Clusters e outro através de Análise de Componentes Principais. Essas variáveis foram introduzidas para agrupar passageiros conforme padrões de consumo das amenidades. A decisão de criar essas variáveis baseou-se na suposição de que padrões de gastos poderiam estar associados à probabilidade de transporte.

4. *Outras Variáveis de Categorização*: HomePlanet, CryoSleep, Destination e VIP foram convertidas em variáveis *dummy* (através de *one -hot encoding*). A categorização dessas variáveis visa transformar informações categóricas em um formato mais adequado para análise por algoritmos de aprendizado de máquina. A decisão de criar variáveis dummy baseou-se na necessidade de representar categorias de forma numérica.

Essas decisões de engenharia de recursos foram orientadas pela intuição sobre os possíveis impactos dessas características na probabilidade de transporte para outra dimensão. O objetivo é melhorar a capacidade dos modelos preditivos em capturar padrões complexos e nuances nos dados.

Durante a fase de pré-processamento, também foram aplicadas algumas transformações nas variáveis para garantir que estivessem em formatos adequados para análise por algoritmos de aprendizado de máquina.

Normalização: garante que as variáveis numéricas estejam em uma escala comparável, evitando distorções nos resultados devido a diferenças nas magnitudes.

1. Age: A variável de idade foi normalizada para lidar com possíveis disparidades de escala entre as características numéricas.

2. Amenidades: As amenidades foram primeiro tratadas através de uma escala logarítmica e posteriormente normalizadas.

Remoção de Variáveis Não Relevantes: Durante o processo de modelagem do Spaceship Titanic, identificamos variáveis que não contribuíam de forma significativa para a previsão. O objetivo dessa prática é simplificar o modelo, mantendo apenas as características mais relevantes. Ao analisar as informações fornecidas, notamos que algumas variáveis, como o nome do passageiro, podem não desempenhar um papel crucial na determinação se um passageiro foi transportado para outra dimensão. Portanto, optamos por remover tais variáveis para reduzir a dimensionalidade do conjunto de dados.

Divisão de Dados

Os dados foram divididos em conjuntos de treino e teste seguindo a abordagem padrão para modelagem de aprendizado de máquina. O arquivo "train.csv" foi dividido em dois conjuntos: 70% dos dados foram destinados ao treinamento do modelo, enquanto os 30% restantes foram reservados para avaliação do desempenho, verificando a capacidade de generalização do modelo para novos dados. Essa escolha visa assegurar que o modelo seja treinado com dados suficientes para aprender padrões significativos.

É importante destacar que, conforme o roteiro de pesquisa da disciplina, uma fase de validação cruzada será realizada ao final da parte 2, no caderno Jupyter, como parte integral do processo de avaliação. O conjunto "test.csv" não contém resultados e foi mantido para a aplicação final do modelo, que não faz parte do escopo deste trabalho.

Essa prática visa evitar o sobreajuste, onde o modelo se adapta excessivamente aos dados de treino específicos e falha ao lidar com novos dados não vistos durante o treinamento. Portanto, a divisão cuidadosa dos dados é crucial para garantir a validade e a eficácia do modelo preditivo para a Spaceship Titanic.

Foram escolhidos três algoritmos de aprendizado de máquina para construir os modelos: Random Forest e Logistic Regression e Support Vector Machine. Além

de Neural Network (da biblioteca scikit-learn) e XGBoost, apenas usados para ilustrar.

A escolha desses algoritmos foi baseada em considerações específicas para o problema abordado. O Random Forest é conhecido por sua eficácia em lidar com conjuntos de dados complexos, apresentando boa performance mesmo sem um ajuste fino extensivo. A Logistic Regression, por sua vez, é apropriada para problemas de classificação binária, o que se alinha com a natureza da competição Kaggle Spaceship Titanic, onde o objetivo é prever se os passageiros foram transportados para outra dimensão ou não. O Support Vector Machine foi escolhido pela sua capacidade de lidar com problemas de classificação e regressão complexos.

A validação cruzada foi realizada para avaliar o desempenho do modelo de Regressão Logística no conjunto de dados. A técnica utilizada foi a validação cruzada *k-fold* com 25 *folds*. Inicialmente, os dados foram padronizados usando o *StandardScaler*. O modelo de Regressão Logística foi criado com um aumento no número máximo de iterações (max_iter=10000) para garantir a convergência do modelo. Em seguida, a validação cruzada foi realizada usando métricas de precisão, recall e F1-score. As pontuações médias dessas métricas foram calculadas a partir dos resultados da validação cruzada.

Ferramentas e Bibliotecas

- Python: Linguagem de programação principal utilizada para análise de dados e implementação de modelos.
- Scikit-learn: Biblioteca em Python para aprendizado de máquina e análise de dados, utilizada para construir e avaliar modelos.
- Pandas: Biblioteca para manipulação e análise de dados.
- NumPy: Biblioteca para operações matemáticas em arrays e matrizes.
- Matplotlib e Seaborn: Bibliotecas para visualização de dados.
- Jupyter Notebooks: Ambiente interativo utilizado para desenvolvimento e documentação do código.
- Outras bibliotecas que foram utilizadas menos extensivamente podem ser obtidas de forma mais completa no início do caderno jupyter.

As análises e modelagens foram executadas em um ambiente computacional padrão Windows. O hardware utilizado foi um notebook processador Intel Core i5, do ano de 2012 e 8 GB de RAM.

Resultados e Discussão

O *dataset* possuía muitos dados faltantes (Gráfico 1) e foram utilizadas várias técnicas para imputar os dados. A princípio se verificou certos padrões, como por exemplo que menores de 13 anos de idade não possuíam despesas ou que estavam em sono criogênico. Isto facilitou suprir estas informações. Quanto a idade, verificamos a existência de um número desproporcional de idades 0 e 1. O que nos fez suspeitar que uma parte dos dados faltantes estava registrada com estes valores. localizamos passageiros que estavam sozinhos com idades de 1 ou 0 anos e reclassificamos como NaN. Usamos a média para suprir as idades faltantes. Tentamos diferentes técnicas para preencher dados faltantes nas diversas categorias, destacando o uso de KNeighborsClassifier e LogisticRegression para planeta de origem e destino, obtendo precisão de 65% em média na base de teste. Nomes e Números de Cabines foram gerados aleatoriamente, pois não havia muito como obter um resultado melhor e nomes seriam descartados na fase posterior.

Validação Cruzada

A Média de Precisão representa a pontuação média de precisão calculada em todos os folds durante a validação cruzada. A precisão mede a proporção de observações positivas corretamente previstas para o total de positivos previstos. Nesse contexto, significa que, em média, cerca de 77,52% das instâncias previstas como positivas pelo modelo durante a validação cruzada eram realmente positivas.

A Média de Recall representa a pontuação média de recall calculada em todos os *folds* durante a validação cruzada. O recall (também conhecido como sensibilidade ou taxa de verdadeiros positivos) mede a proporção de observações positivas corretamente previstas para todos os positivos reais. Uma média de recall de cerca de 79,46% indica que, em média, o modelo capturou aproximadamente 79,46% de todas as instâncias positivas reais durante a validação cruzada.

A Média de Pontuação F1 representa a pontuação média de F1 calculada em todos os *folds* durante a validação cruzada. A pontuação F1 é a média harmônica de precisão e recall. Ela proporciona um equilíbrio entre precisão e recall. Uma média

de pontuação F1 de cerca de 78,41% sugere que, em média, o modelo alcançou um equilíbrio entre precisão e recall, com 78,41% sendo a média harmônica de precisão e recall.

A Precisão no Teste representa a pontuação de precisão calculada no conjunto de dados de teste. Isso significa que, de todas as instâncias previstas como positivas pelo modelo nos dados de teste, aproximadamente 75,99% eram verdadeiras instâncias positivas.

O Recall no Teste representa a pontuação de recall calculada no conjunto de dados de teste. Isso indica que cerca de 80,36% de todas as instâncias positivas reais nos dados de teste foram corretamente identificadas pelo modelo.

A Pontuação F1 no Teste representa a pontuação F1 calculada no conjunto de dados de teste. Uma pontuação F1 no teste de cerca de 78,11% sugere que o modelo alcançou um equilíbrio entre precisão e recall, com 78,11% sendo a média harmônica de precisão e recall nos dados de teste.

Tabela 1: Matriz de Confusão

Matriz de Confusão	
76,64%	23,36%
20,59%	79,40%

A análise da matriz de confusão (Tabela 1), que optamos em apresentar em termos percentuais, revela um desempenho equilibrado do modelo. A taxa de verdadeiros positivos atinge 79,34%, indicando uma boa capacidade de identificar corretamente casos positivos, enquanto a taxa de verdadeiros negativos é de 76,57%, refletindo uma precisão sólida na previsão de casos negativos. No entanto, observa-se uma taxa de falsos positivos de 23,43% e a taxa de falsos negativos é de 20,66%, sugerindo uma proporção relativamente alta de previsões incorretas, porém dentro do esperado para este *dataset*.

Decomposição em Componentes (PCA) e Clustering

Aplicamos a Decomposição em Componentes Principais às variáveis numéricas, especificamente à idade e às despesas. Do gráfico 10 (abaixo), podemos inferir que o primeiro componente pode capturar 35% da variância e o segundo 20%. Isso significa que mais da metade da variância pode ser expressa por apenas duas variáveis.

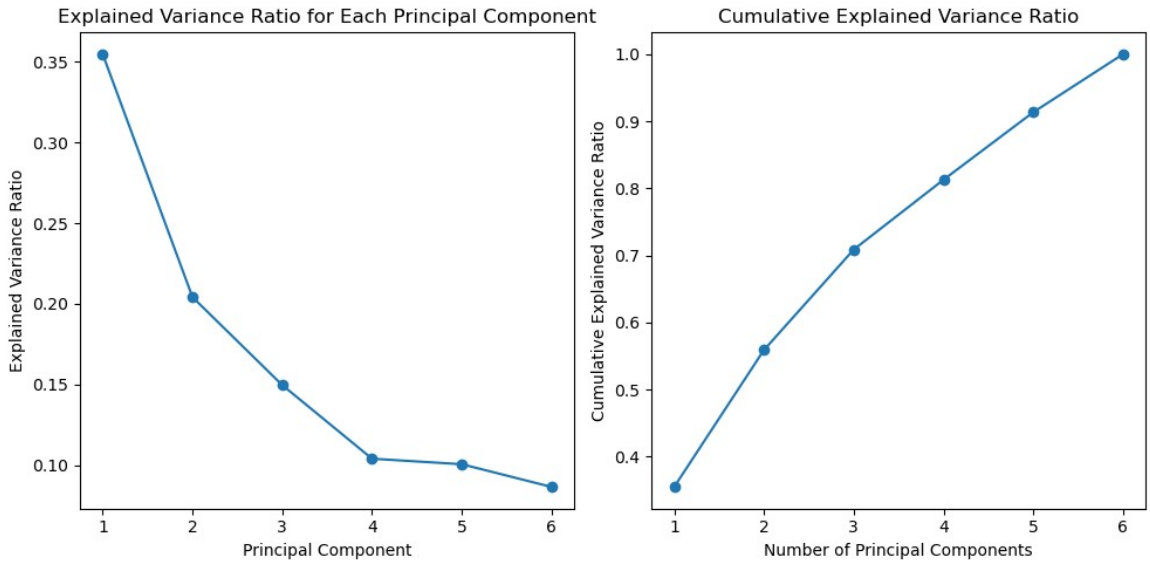

Gráfico 10: Decomposição em Componentes Principais

Para testar a relevância destas componentes aplicamos a regressão logística para avaliar a quantidade de informação que de fato cada componente traz para predição da coluna “Transported” da nossa base de treino.

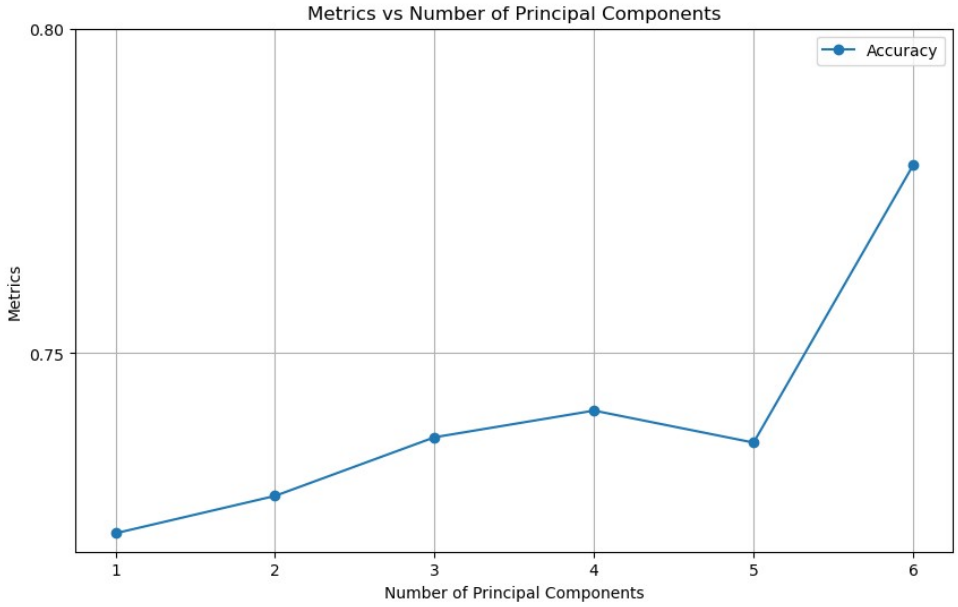

Gráfico 11: Número de Componentes Principais contra Capacidade Preditiva

No modelo inicial (Gráfico 11) com 2 componentes principais, a precisão atingiu aproximadamente 72,8%, indicando que essa representação simplificada captura efetivamente uma parte significativa das informações relevantes para prever o status de "Transportado". A adição de um terceiro componente principal levou a uma melhoria leve, alcançando cerca de 73,7%, indicando sua valiosa contribuição

para o poder preditivo. A inclusão adicional de um quarto componente principal resultou em uma precisão de aproximadamente 74,1%, sugerindo insights aprimorados e melhor desempenho preditivo.

Surpreendentemente, a inclusão de um quinto componente principal resultou em uma ligeira queda na precisão para cerca de 73,6%. Essa diminuição inesperada pode ser atribuída a ruído ou padrões menos informativos capturados por esse componente específico nos dados.

A maior precisão entre as configurações testadas foi alcançada ao usar seis componentes principais, atingindo aproximadamente 77,9%. Isso indica que a inclusão do sexto componente principal aprimora a capacidade do modelo de discernir padrões relacionados ao status de *Transported*.

A precisão da previsão de controle, sem utilizar qualquer análise de PCA, foi de 77,1%. Isso sugere que o desempenho do modelo com os seis principais componentes é (como esperado) comparável à precisão alcançada sem a redução de dimensionalidade.

A seguir, então, procuramos quais as componentes mais relevantes para obter esse resultado. As melhores 5 combinações de componentes principais foram identificadas iterando por todas as combinações possíveis, classificadas de acordo com suas precisões correspondentes. Notavelmente, a primeira combinação, com uma precisão de 77,49%, consiste em PC1, PC2, PC4 e PC6, alcançando a mesma precisão que a segunda combinação que substitui PC2 por PC3.

O próximo exercício foi o de obter agrupamentos, *Clustering*. Utilizamos o método gráfico do cotovelo para encontrar o valor ótimo de k. O valor ótimo de K é o ponto em que o gráfico forma um cotovelo (Gráfico 12). A seguir produzimos um gráfico de todas as categorias por todas as categorias para observar a formação de centroides (Gráfico 13). O gráfico sugere que o algoritmo identificou três clusters principais, com os clusters verde e amarelo apresentando padrões mais claros frequentemente em direções opostas.

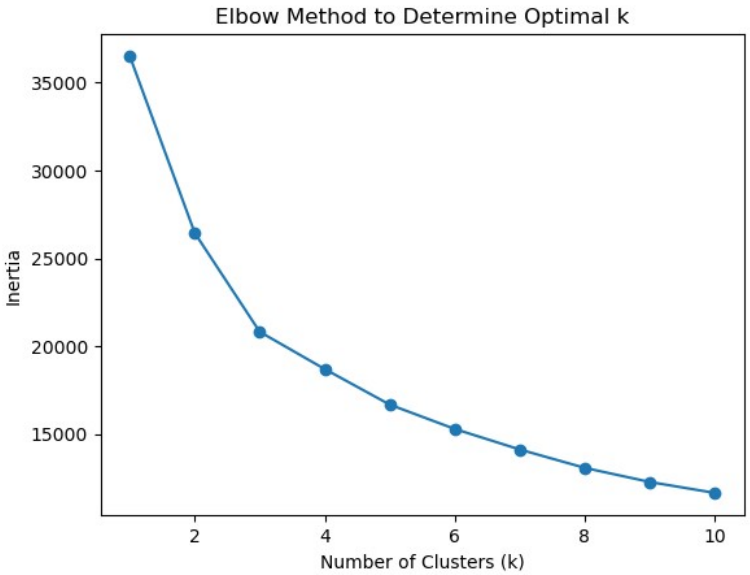

Gráfico 12: Método do Cotovelo

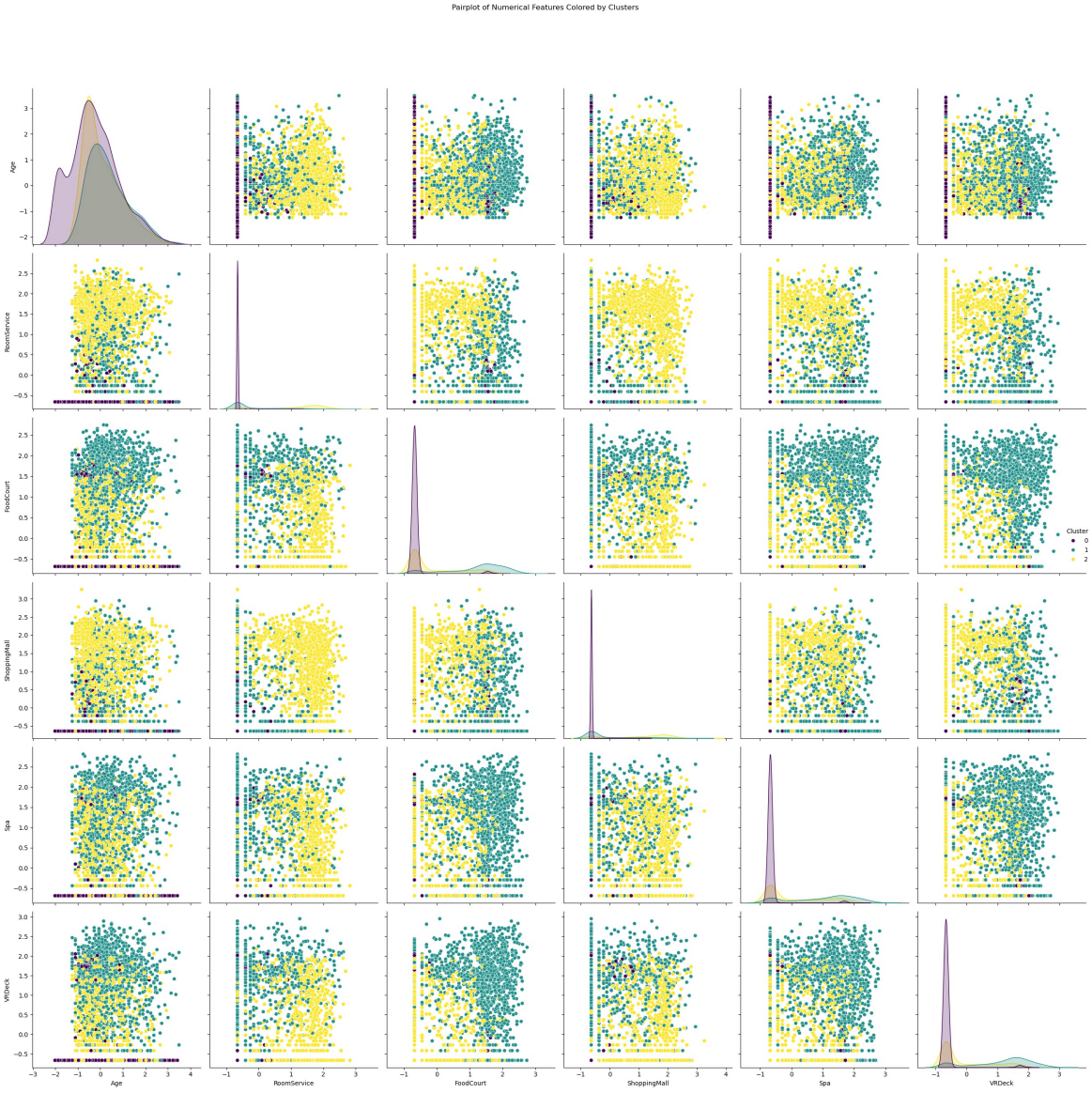

Gráfico 13: Visualização de clusters através das diferentes combinações de categorias

Após aplicarmos o mesmo teste de regressão logística utilizado anteriormente com as componentes principais, observamos que a abordagem de clusterização apresentou um desempenho superior em comparação com a análise de PCA, especialmente ao empregar menos componentes. É notável que o modelo de clusterização com k=3 tenha alcançado uma precisão significativa de 74,77%, superando o modelo PCA com três componentes principais que atingiu 73,71% de precisão. Esses resultados sugerem que os padrões inerentes capturados pela clusterização de características numéricas são mais eficazes na previsão da coluna *Transported* em comparação com os componentes principais derivados do PCA.

Uma vantagem adicional dessa abordagem é a conversão de valores numéricos para booleanos, o que simplifica a próxima etapa ao permitir que prossigamos apenas com valores booleanos, se necessário. Este aspecto destaca a flexibilidade e a eficácia da abordagem de clusterização em nosso contexto específico.

Por fim, realizamos a quarta parte do trabalho que consiste em aplicar mais de um modelo. Escolhemos a Regressão Logística, que tem sido nossa técnica de controle através deste estudo, Random Forest e Support Vector Machine. Adicionalmente incluímos Neural Network e XGBoost com a finalidade de melhor ilustrar a facilidade de se incluir modelos no estudo.

A Regressão Logística, embora nomeada assim, é um algoritmo de classificação. É empregada quando a variável dependente é binária, modelando a probabilidade de uma instância pertencer a uma categoria específica. Ao usar a função logística para transformar a saída de uma equação linear em uma pontuação de probabilidade, a Regressão Logística facilita a classificação com base em um limiar.

O Random Forest é uma abordagem de aprendizado em conjunto que constrói múltiplas árvores de decisão durante o treinamento. Utilizando subconjuntos aleatórios dos dados e características para cada divisão, esse método busca aprimorar a generalização do modelo, evitando o *overfitting*. Ao combinar as previsões de cada árvore, o Random Forest proporciona resultados mais robustos.

As Máquinas de Vetores de Suporte (SVMs) são modelos poderosos para aprendizado supervisionado, aplicados tanto em classificação quanto em regressão. No contexto da classificação, as SVMs buscam um hiperplano ideal para separar os dados em classes, identificando vetores de suporte próximos à fronteira de decisão.

Eficientes em espaços de alta dimensão, as SVMs são especialmente úteis quando as classes não são linearmente separáveis.

Os resultados que obtivemos apontam para uma performance equivalente entre os três modelos em todas as métricas. Os modelos considerados abaixo incluem Random Forest, Regressão Logística, Máquina de Vetores de Suporte (SVM), Rede Neural e XGBoost, embora os dois últimos tenham sido adicionados apenas de forma ilustrativa. O conjunto de dados foi pré-processado, as características foram selecionadas, e os modelos foram treinados e avaliados quanto ao seu desempenho preditivo, conforme resultados abaixo:

Desempenho dos Modelos

1. Random Forest
    - Acurácia: 78,26%
    - Precisão: 77% (Treino), 80% (Teste)
    - Recall: 80% (Treino), 77% (Teste)
    - F1-Score: 78% (Treino), 78% (Teste)
2. Regressão Logística
    - Acurácia: 76,84%
    - Precisão: 78% (Treino), 76% (Teste)
    - Recall: 74% (Treino), 80% (Teste)
    - F1-Score: 76% (Treino), 78% (Teste)
3. Máquina de Vetores de Suporte (SVM)
    - Acurácia: 77,38%
    - Precisão: 77% (Treino), 78% (Teste)
    - Recall: 77% (Treino), 78% (Teste)
    - F1-Score: 77% (Treino), 78% (Teste)
4. Rede Neural
    - Acurácia: 77,49%
    - Precisão: 76% (Treino), 79% (Teste)
    - Recall: 79% (Treino), 76% (Teste)
    - F1-Score: 78% (Treino), 77% (Teste)
5. XGBoost
    - Acurácia: 77,76%

- Precisão: 78% (Treino), 77% (Teste)
- Recall: 76% (Treino), 79% (Teste)
- F1-Score: 77% (Treino), 78% (Teste)

Random Forest apresentou a maior acurácia entre os modelos, porém a vantagem é marginal enquanto melhorias podem ser feitos para substancialmente se obterem melhores resultados no tratamento dos dados, após este estudo preliminar. Particularmente o tratamento de outliers escolhido, que estabelece um teto para despesas máximas no 99º percentil, parece reduzir em meio porcento a acurácia dos modelos, exceto em XGBoost que, ao contrário, melhora com esse tratamento.

Conclusão

Ao longo deste estudo, exploramos diversas abordagens de análise, incluindo a aplicação de técnicas de PCA, *clustering* e modelos de aprendizado de máquina, como Random Forest, Regressão Logística e Máquinas de Vetores de Suporte (SVMs).

Percebemos a importância de estudar o comportamento das variáveis para suprir dados faltantes e mesmo criar hipóteses sobre a causa de certas características dos dados.

Fomos expostos a uma enormidade de situações que exigiram pesquisa e permitiram a aplicação de conhecimentos obtidos no decorrer dos dois cursos de aprendizado de máquinas que participamos este ano.

Em um aspecto mais técnico, observamos que a clusterização, com um valor ótimo de k=3, superou a análise de PCA em termos de previsão da variável *Transported*, alcançando uma precisão notável de 74,77%. Esta descoberta ressalta a eficácia da clusterização de características numéricas na captura de padrões subjacentes ao fenômeno estudado.

Os valores de precisão e acurácia para esta competição são considerados bons a partir de 80%. Neste sentido, consideramos os resultados aqui obtidos, na faixa de 78%, bastante adequados posto que não estamos usando todos os dados para o treino. Estes resultados intermediários, também, apontam para uma performance equivalente entre os modelos em todas as métricas para essa base de dados, com uma certa vantagem para Random Forest. A próxima etapa será

entender quais variáveis podem se beneficiar de um tratamento diferente e treinar o modelo com todos os dados de treino para então aplicar aos dados teste fornecidos pelo *Kaggle* para submissão à competição.

Referências

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning : with Applications in R**. New York :Springer, 2013. 607p.