**Trabalho final de Mineração de dados - Dataset Reservas de Hotel.**

**Introdução**

O relatório final em questão tem como objetivo analisar um conjunto de dados referentes às reservas feitas por clientes da rede 'INN Hotels', incluindo informações sobre preço, tempo de estadia e pedidos especiais.

O foco da análise será identificar quais variáveis têm mais influência na previsão de cancelamentos de diárias, individualmente ou em relação a outras variáveis, e criar um modelo de classificação para futuras reservas.

A métrica de precisão será utilizada como referência para avaliar a performance dos modelos de classificação, priorizando a minimização do erro de falso positivo, ou seja, evitar previsões incorretas de cancelamentos de diárias.

**Análise exploratória dos dados**

**Dicionário de dados**:

Booking\_ID: identificador único de cada reserva (ordinal)

no\_of\_adults: número de adultos (discreta)

no\_of\_children: número de crianças (discreta)

no\_of\_weekend\_nights: número de noites em fim de semana em que o hóspede ficou ou se hospedou para ficar no hotel (discreta)

no\_of\_week\_nights: número de noites durante a semana em que o hóspede ficou ou se hospedou para ficar no hotel (discreta)

type\_of\_meal\_plan: tipo de plano de alimentação (nominal)

required\_car\_parking\_space: requisitou vaga de estacionamento? (nominal)

room\_type\_reserved: tipo de acomodação reservada (nominal)

lead\_time: número de dias entre a data de reserva e a data de chegada (contínua)

arrival\_year: ano de chegada (ordinal)

arrival\_month: mês de chegada (ordinal)

arrival\_date: dia de chegada (ordinal)

market\_segment\_type: segmento de mercado (nominal)

repeated\_guest: é um cliente reincidente? (nominal)

no\_of\_previous\_cancellations: número de reservas anteriores que foram canceladas pelo hóspede (discreta)

no\_of\_previous\_bookings\_not\_canceled: número de reservas anteriores que não foram canceladas pelo hóspede (discreta)

avg\_price\_per\_room: preço médio por dia de reserva, onde o preço dos quartos são dinâmicos (contínua)

no\_of\_special\_requests: número total de pedidos especiais feitos pelo hóspede, como andar alto, vista, etc (discreta)

booking\_status: variável target que indica se a reserva foi cancelada ou não

**Pré Processamento:**

Após a análise do conjunto de dados, conclui-se que este possui 19 colunas e 36275 linhas, sendo que a coluna "Booking\_ID" não agrega valor ao processo classificatório e será abandonada. Não foram encontrados dados nulos, mas existem dados ausentes em algumas variáveis.

A variável "booking\_status", objeto principal de análise, será convertida para o tipo booleano. Além disso, as variáveis 'required\_car\_parking\_space' e 'repeated\_guest' também serão convertidas para booleano.

As variáveis categóricas 'type\_of\_meal\_plan', 'room\_type\_reserved' e 'market\_segment\_type' serão convertidas para o tipo int64 para possibilitar uma análise mais ampla. O dicionário de dados e a análise de valores únicos foram utilizados para orientar as conversões de tipos.

**Limpeza inicial**

Primeiramente, a coluna 'Booking\_ID' foi removida do dataframe. Em seguida, foram definidos dois dicionários, um para converter valores booleanos e outro para converter valores categóricos para inteiros.

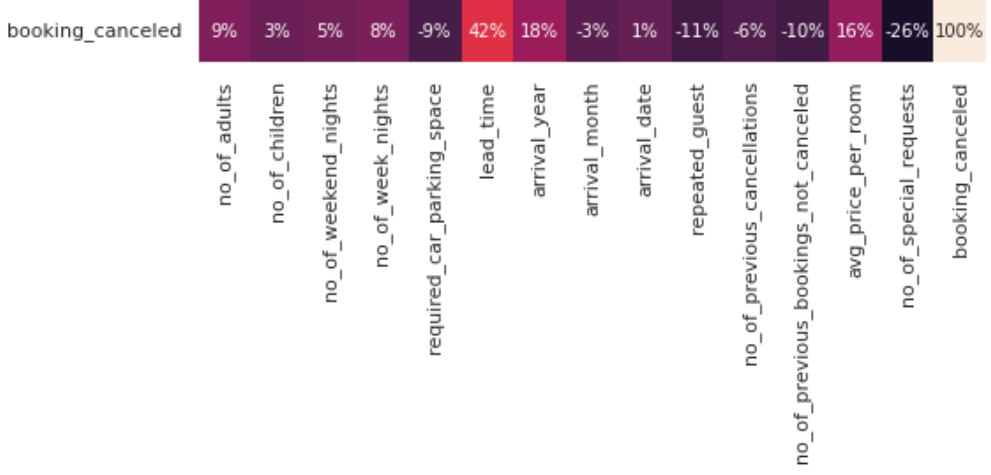
Foram aplicadas as conversões de tipos para as variáveis utilizando os dicionários mencionados para assim categorizar as variáveis 'required\_car\_parking\_space', 'repeated\_guest' e 'booking\_status', enquanto as variáveis categóricas 'market\_segment\_type', 'type\_of\_meal\_plan' e 'room\_type\_reserved' foram definidas como tipo “Category”. Por fim, a coluna 'booking\_status' foi renomeada para 'booking\_canceled'.

**Análise univariada**

Com base nas informações apresentadas, pode-se afirmar que o conjunto de dados é considerado bem comportado. Isso porque, segundo os pontos apresentados, pouca manipulação será exigida em um primeiro momento, o que pode indicar que os dados estão organizados e estruturados de maneira clara e coesa.

Além disso, o fato de não ser necessário preencher dados ausentes indica que o conjunto de dados está completo, o que é um ponto positivo na análise. Outro aspecto importante é a presença de poucos outliers, que são coerentes com o contexto, o que sugere que esses valores discrepantes não comprometem a análise do conjunto de dados.  
 **Análise multivariada**

Por meio da análise do heatmap, observa-se uma forte correlação (93%) entre as variáveis 'no\_of\_previous\_bookings\_not\_canceled' e 'repeated\_guest', o que sugere que hóspedes recorrentes (1) tendem a não cancelar suas reservas. Essa relação pode ser justificada por diversos fatores, como a existência de benefícios atrelados à recorrência, a sensação de pertencimento ou identificação com o local, entre outros.

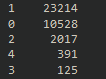
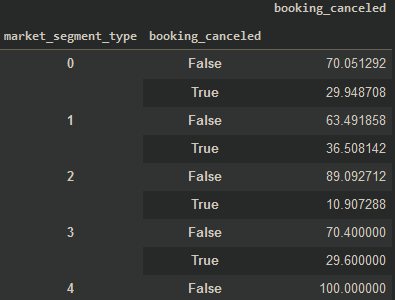


Além disso, a variável 'repeated\_guest' também apresenta uma alta correlação (60%) com 'no\_of\_previous\_cancellations', o que sugere que somente hóspedes que já estiveram no estabelecimento anteriormente têm registros de cancelamentos de reservas.

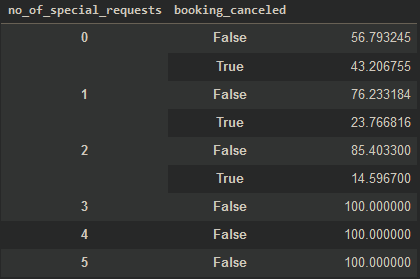
Outra correlação que merece destaque é entre as variáveis 'lead\_time' e 'booking\_canceled' (42%), indicando que quanto maior o intervalo de tempo entre a reserva e o check-in, maior a probabilidade de cancelamento da reserva. Reservas de última hora, com base no primeiro quartil, apresentam uma taxa de cancelamento de apenas 11%, o que reforça a tese anterior.

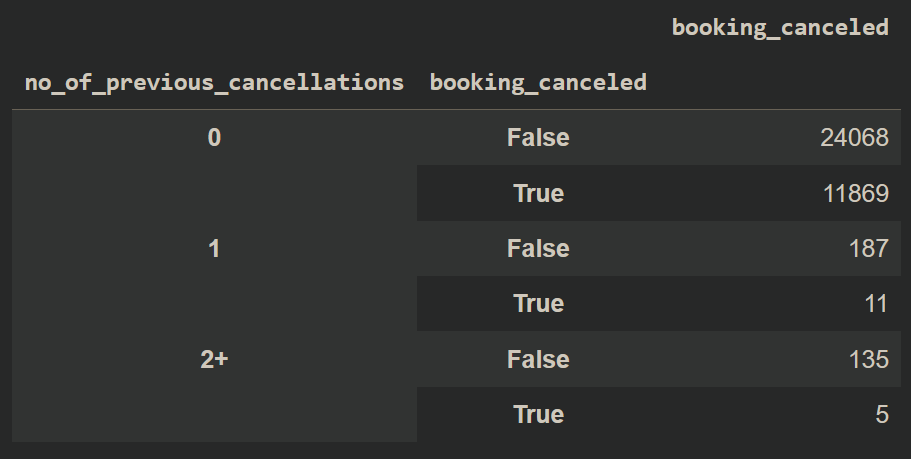
**Correlação com variável Target**

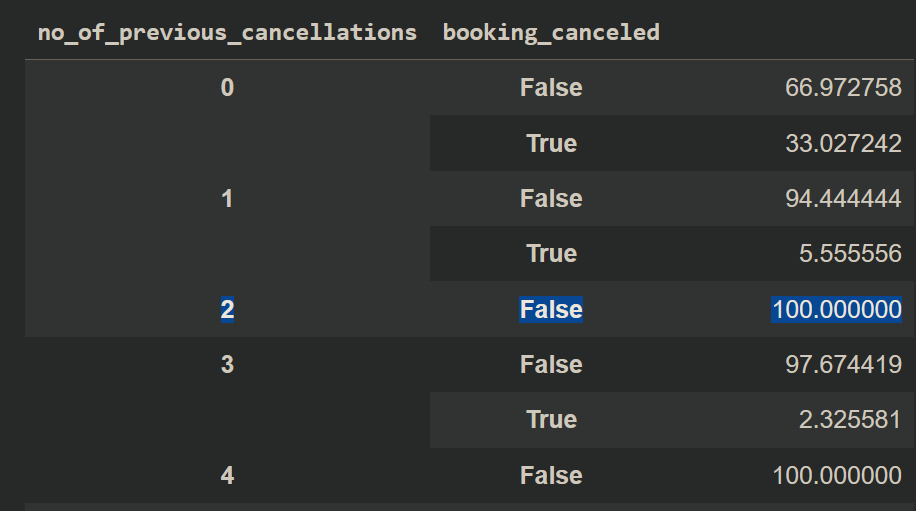
Ao analisar a relação entre o tipo de segmento de mercado e o cancelamento de reservas, destaca-se que nenhum cliente do segmento 4 cancelou sua reserva. Isso sugere que é possível aplicar um filtro no dataset, considerando que os clientes desse grupo não tendem a cancelar suas reservas. A presença de um número significativo de clientes nesse segmento reforça essa conclusão.



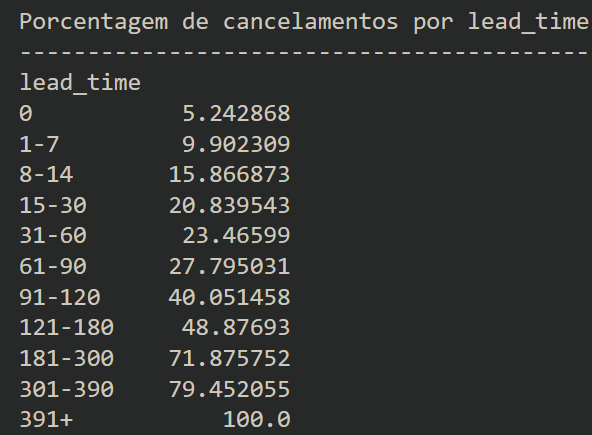
É evidente a partir dos dados que as pessoas que fazem pedidos especiais são menos propensas a cancelar suas solicitações. Além disso, observamos que, quando há mais de três pedidos especiais, há uma certeza de 100%, de acordo com o conjunto de dados, de que o cliente não irá cancelar. Isso é apoiado pelo número de clientes presentes nesses três segmentos, o que sugere que podemos filtrar o conjunto de dados e considerar que qualquer cliente com mais de três pedidos especiais não irá cancelar.



É uma possível conclusão que indivíduos com no\_of\_previous\_cancellations > 1 não tendem a cancelar agendamentos, embora a amostragem dessas pessoas não seja significativa. 

No entanto, aqueles com no\_of\_previous\_cancellations = 1 têm uma taxa de cancelamento de apenas 5,6% (em comparação com 33% para no\_of\_previous\_cancellations = 0) e uma amostragem considerável, o que sugere que a hipótese forte.. 

Dependendo do comportamento do algoritmo utilizado, podemos filtrar os dados para classificar no\_of\_previous\_cancellations > 1 como não cancelamento, sabendo que podemos ter um erro de cerca de 6%, o que pode ser menor do que o erro que o algoritmo é capaz de inferir.

É notável que a taxa de cancelamento cresce progressivamente à medida que o lead\_time aumenta, atingindo 100% acima de 390 dias. No modelo final, é possível filtrar os dados com lead\_time>390 e afirmar com certeza que esses agendamentos serão cancelados.

Em uma situação real, essa informação seria suficiente para adotar medidas preventivas, como cobrar uma taxa para agendamentos com mais de ~181 dias (quando a taxa de cancelamento aumenta de 38% para 71%).

**Modelos de predição utilizados**

**Separação de dados**

O dataset tratado foi dividido em duas técnicas diferentes, apenas Houdout e Crossvalidation + Houdout. Este último é feito a fim de evitar enviesar os dados de treinamento com a resposta do dataframe de teste, para esse fim, foi utilizado o número canônico de “folds” igual a 10 . O dataset extraído para o Houdout foi de 20% das linhas, já que havia mais de 36 mil linhas no total.

**Modelos aplicados:**

**KNN**

Para o algoritmo KNN pela separação de dados houdolt foram usados as seguintes métricas:

n\_neighbors=25, weights='distance', algorithm='auto', leaf\_size=30, p=2, metric='minkowski'

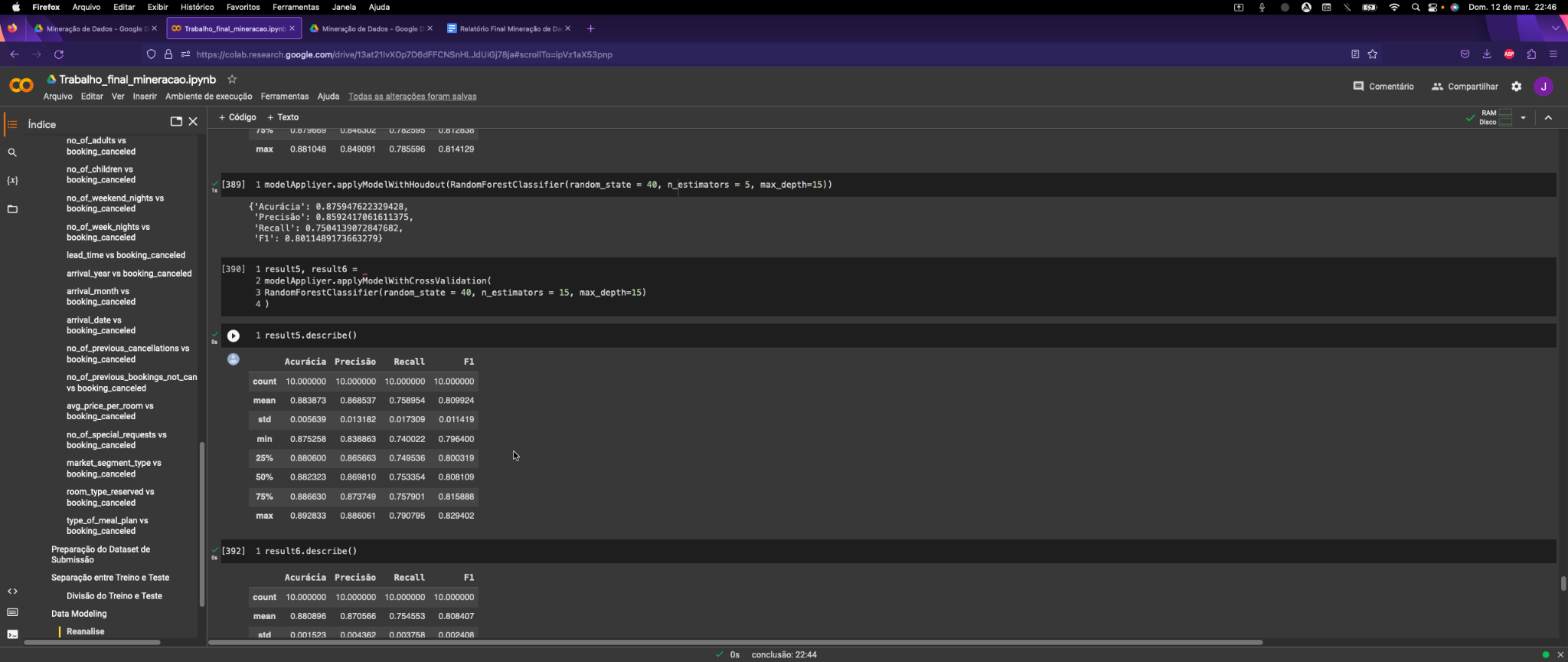
Foi obtido 84,74% de precisão, o que foi considerado com um bom resultado, já com separação do Crossvalidation a média da precisão foi de 84,53% e as métricas foram as mesmas. A fim de alcançar um melhor resultado foi variado o parâmetro neighbors, anteriormente com o número desse parâmetro em 3, 10 e 15 obtivemos uma precisão inferior.

**Decision Tree**

Decision Tree foi utilizado apenas para demonstração já que é apenas um subcaso do Random Forest, o melhor modelo obtido teve uma precisão de 82%, foi variado o parâmetro max\_depth com valores até 25, sendo o melhor parâmetro 15.

**Random forest**

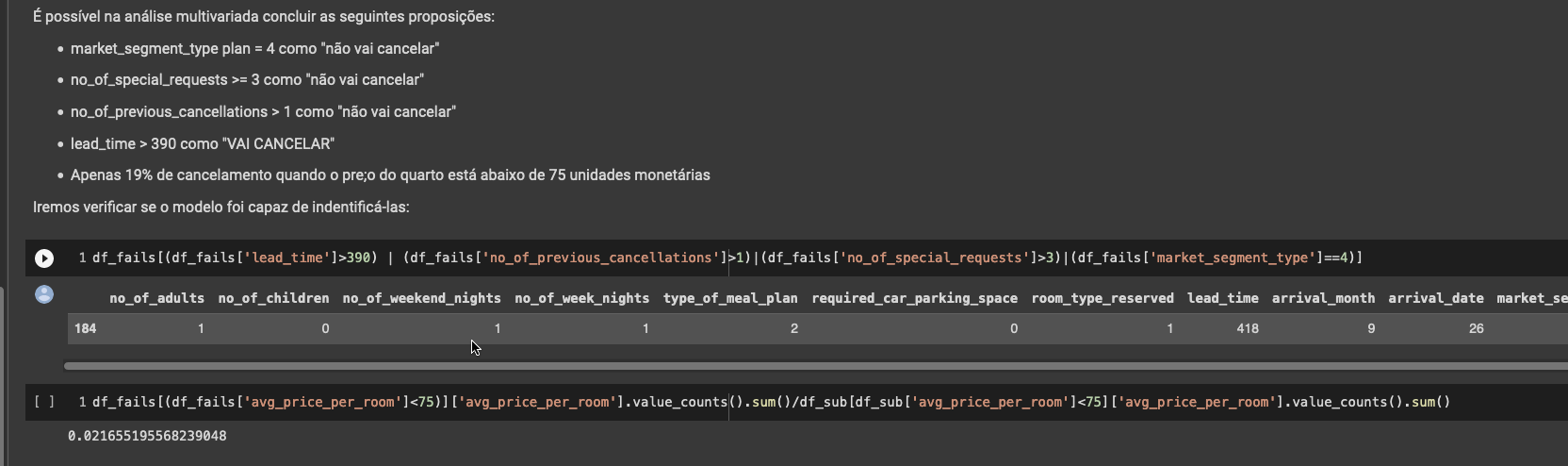
Para o random forest foram usados os seguintes parâmetros:

****

Obtivemos o melhor modelo utilizando CrossValidation, no fold de número 4 que apresentou uma precisão de 87,90%, sendo o melhor de todos os modelos testados. Foi variado o número de estimadores de 5 a 20, sendo 15 o melhor resultado.

**Onde o modelo erra mais?**

Conclui-se através dos testes realizados que o modelo conseguiu generalizar as partes relevantes dos dados, embora tenha tido dificuldade em compreender os dados mais complexos.



Entretanto, considerando o baseline de 67% (que representa a porcentagem de dados que não tiveram a reserva cancelada no conjunto ), pode-se afirmar que o modelo obteve um bom desempenho.

**Resultados do melhor modelo**

Conforme a imagem abaixo, podemos ver que o fold de número 4 obteve o melhor resultado dentro do Cross Validation + Houtdout utilizando Random Florest.

Neste modelo em especial, obtivemos a precisão de 0.879. Apesar de ser o melhor modelo, podemos observar que as variações dos demais modelos é baixa, corroborando que temos um modelo que realmente condiz com a aplicação. Além disso, os resultados dos demais modelos (decision tree e KNN) estão dentro de uma margem de 10% de precisão, indicando que todos os modelos poderiam ser utilizados de forma coerente.