Modelo de Predição de Cancelamento de Reservas em Hotéis

SCC0230 Inteligência Artificial

ICMC USP SÃO CARLOS

Descrição da Problemática:

- Novos métodos de reserva de estadias em hotéis facilitam a experiência do usuário, porém podem trazer novos problemas para os hotéis.
- A comodidade pode fazer com que o cliente trate com menor importância a reserva.
- Reservas via site de domínio próprio ou aplicativos e/ou portais de reserva de terceiros distanciam o cliente da empresa.
- Maior comprometimento nas reservas mais “analógicas”, como via telefone ou até presencialmente.
- Cabe aos hotéis buscarem soluções para amenizar as chances do cliente não honrar com sua reserva, ou ao menos se preparar para isso.
- Busca-se mitigar o ônus dos hotéis utilizando modelos preditivos inteligentes.

Os dados:

- Booking ID
- no_of_adults
- no_of_childrens
- no_of_weekend_nights
- no_of_week_nights
- type_of_meal_plan
- required_car_parking_space
- room_type_reserved
- lead_time
- arrival_year
- arrival_month
- arrival_date
- market_segment_type
- repeated_guest
- no_of_previous_cancellations
- no_of_previous_bookings_not_canceled
- avg_price_per_room
- no_of_special_requests
- booking_status

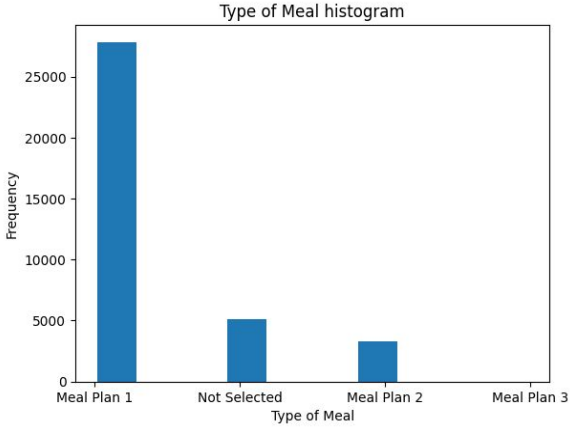
Pré-processamento dos dados:

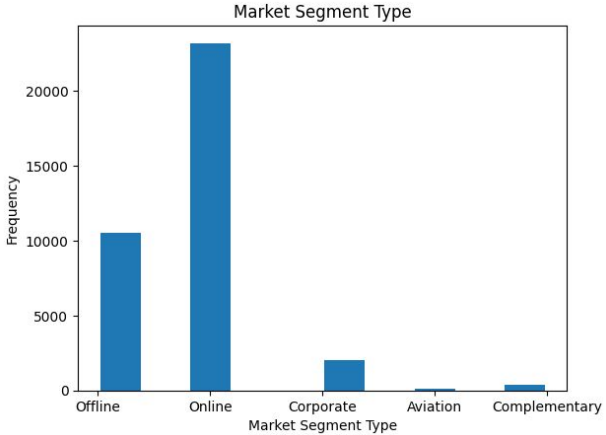
O dataset já se encontrava limpo, sem dados duplicados ou nulos, dispensando a necessidade de pré-processamento dos dados. Houve apenas uma preparação - distinta para cada modelo - antes de cada modelagem.

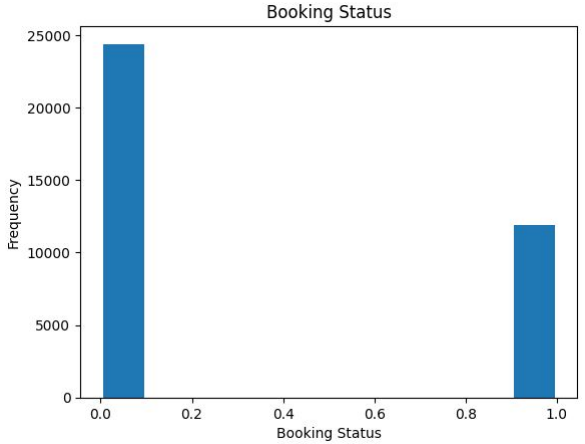
Análise dos dados:

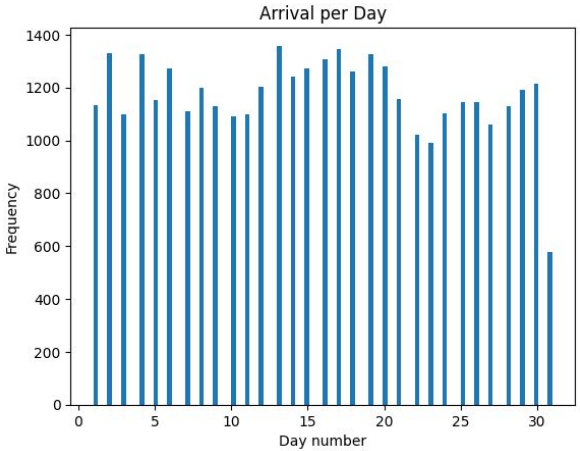
Histograma das variáveis categóricas

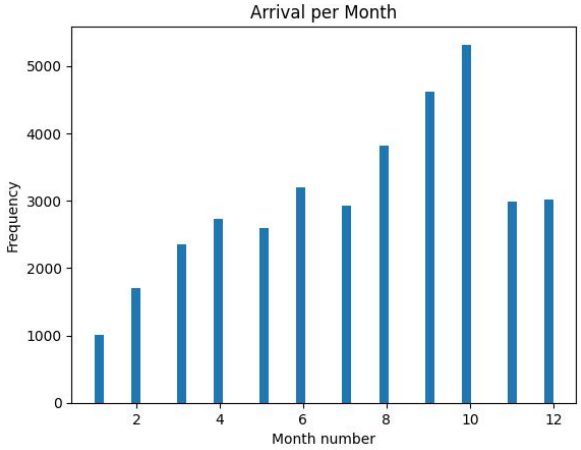
A investigação dessas variáveis visa fornecer insights sobre preferências, comportamentos e características distintas dos clientes. Os resultados dessa análise estão apresentados no slide seguinte.

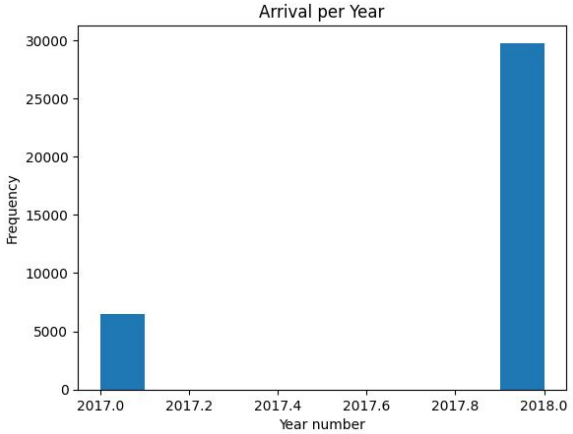

Type of Meal histogram
Frequency
25000
20000
15000
10000
5000
0
Meal Plan 1
Not Selected
Meal Plan 2
Meal Plan 3
Type of Meal


Market Segment Type
Frequency
20000
15000
10000
5000
0
Offline
Online
Corporate
Aviation
Complementary
Market Segment Type


Booking Status
Frequency
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
Booking Status


Arrival per Day
Frequency
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
Day number

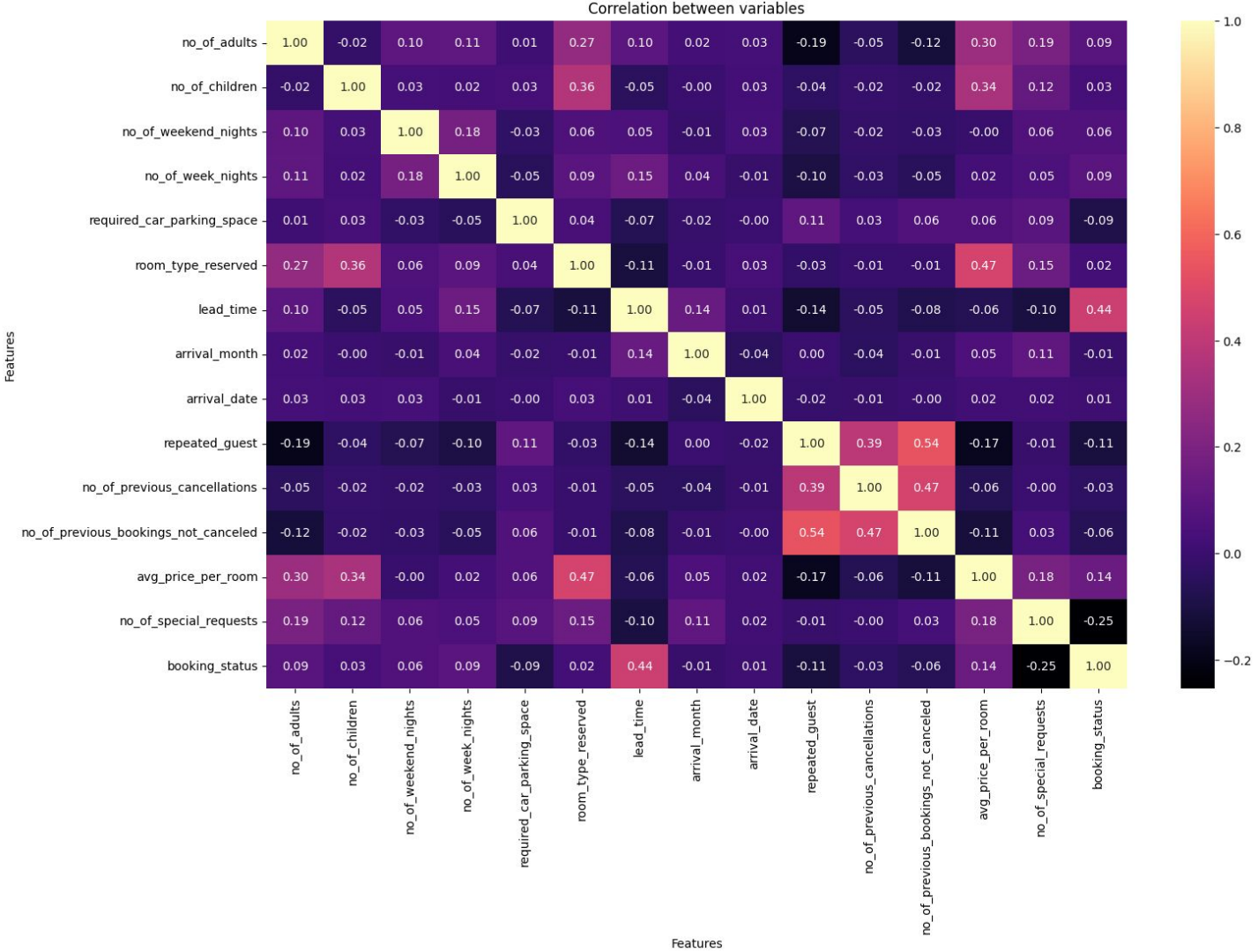

Arrival per Month
Frequency
5000
4000
3000
2000
1000
0
2
4
6
8
10
12
Month number


Arrival per Year
Frequency
30000
25000
20000
15000
10000
5000
0
2017.0
2017.2
2017.4
2017.6
2017.8
2018.0
Year number

Análise dos dados:

Correlação entre as variáveis

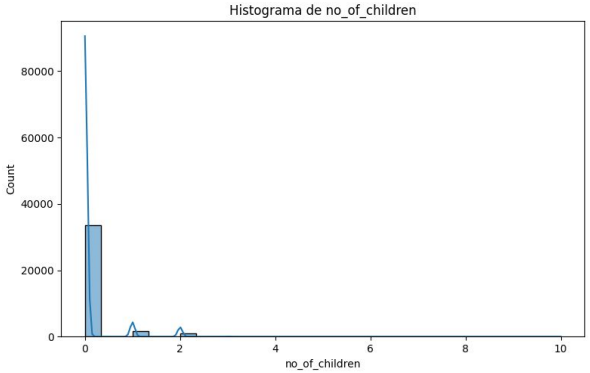
Conduzimos uma análise de correlação com o objetivo de explorar as relações entre as variáveis em nosso conjunto de dados de reservas de hotel, buscando entender as interconexões e padrões subjacentes. Os resultados estão detalhados no slide subsequente.


Correlation between variables
Features
no_of_adults 1.00 -0.02 0.10 0.11 0.01 0.27 0.10 0.02 0.03 -0.19 -0.05 -0.12 0.30 0.19 0.09
no_of_children -0.02 1.00 0.03 0.02 0.03 0.36 -0.05 -0.00 0.03 -0.04 -0.02 -0.02 0.34 0.12 0.03
no_of_weekend_nights 0.10 0.03 1.00 0.18 -0.03 0.06 0.05 -0.01 0.03 -0.07 -0.02 -0.03 -0.00 0.06 0.06
no_of_week_nights 0.11 0.02 0.18 1.00 -0.05 0.09 0.15 0.04 -0.01 -0.10 -0.03 -0.05 0.02 0.05 0.09
required_car_parking_space 0.01 0.03 -0.03 -0.05 1.00 0.04 -0.07 -0.02 -0.00 0.11 0.03 0.06 0.06 0.09 -0.09
room_type_reserved 0.27 0.36 0.06 0.09 0.04 1.00 -0.11 -0.01 0.03 -0.03 -0.01 -0.01 0.47 0.15 0.02
lead_time 0.10 -0.05 0.05 0.15 -0.07 -0.11 1.00 0.14 0.01 -0.14 -0.05 -0.08 -0.06 -0.10 0.44
arrival_month 0.02 -0.00 -0.01 0.04 -0.02 -0.01 0.14 1.00 -0.04 0.00 -0.04 -0.01 0.05 0.11 -0.01
arrival_date 0.03 0.03 0.03 -0.01 -0.00 0.03 0.01 -0.04 1.00 -0.02 -0.01 -0.00 0.02 0.02 0.01
repeated_guest -0.19 -0.04 -0.07 -0.10 0.11 -0.03 -0.14 0.00 -0.02 1.00 0.39 0.54 -0.17 -0.01 -0.11
no_of_previous_cancellations -0.05 -0.02 -0.02 -0.03 0.03 -0.01 -0.05 -0.04 -0.01 0.39 1.00 0.47 -0.06 -0.00 -0.03
no_of_previous_bookings_not_canceled -0.12 -0.02 -0.03 -0.05 0.06 -0.01 -0.08 -0.01 -0.00 0.54 0.47 1.00 -0.11 0.03 -0.06
avg_price_per_room 0.30 0.34 -0.00 0.02 0.06 0.47 -0.06 0.05 0.02 -0.17 -0.06 -0.11 1.00 0.18 0.14
no_of_special_requests 0.19 0.12 0.06 0.05 0.09 0.15 -0.10 0.11 0.02 -0.01 -0.00 0.03 0.18 1.00 -0.25
booking_status 0.09 0.03 0.06 0.09 -0.09 0.02 0.44 -0.01 0.01 -0.11 -0.03 -0.06 0.14 -0.25 1.00
no_of_adults
no_of_children
no_of_weekend_nights
no_of_week_nights
required_car_parking_space
room_type_reserved
lead_time
arrival_month
arrival_date
repeated_guest
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
avg_price_per_room
no_of_special_requests
booking_status
Features
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

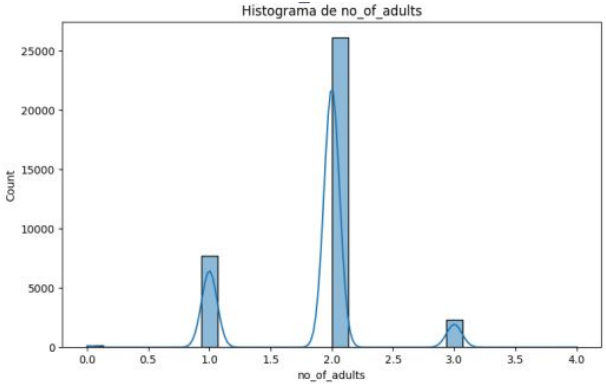
Análise dos dados:

Distribuição das variáveis contínuas

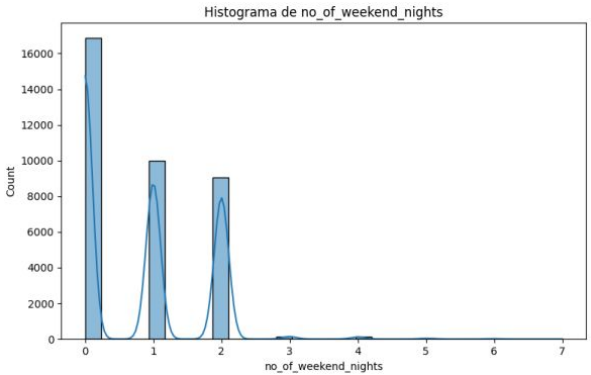
Fizemos uma análise exploratória com o intuito de obter uma visão abrangente das características de algumas das variáveis contínuas em nosso conjunto de dados de reservas de hotel, como apresentado no slide a seguir.



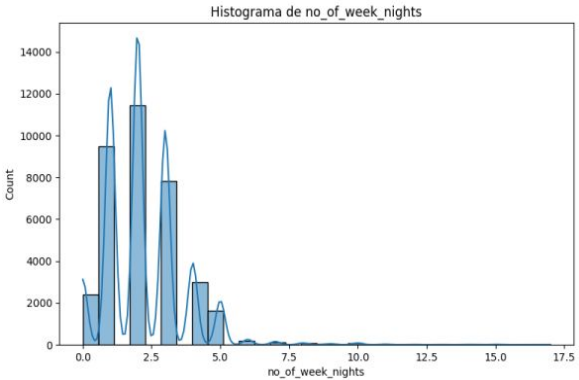
- Examina a distribuição da presença de crianças nas reservas.
- Ajuda a observar a frequência de reservas com ou sem crianças.



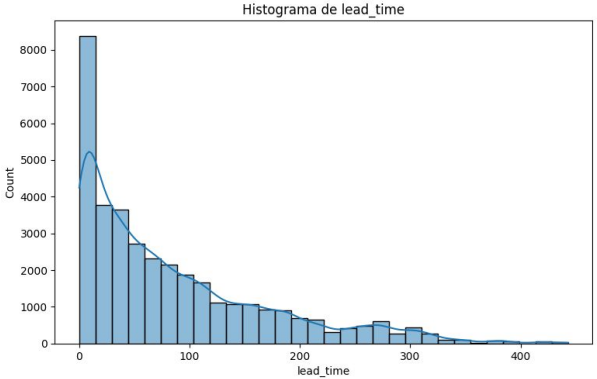
- O histograma revela a distribuição do número de adultos nas reservas.
- Permite identificar se a maioria das reservas é para casais, grupos ou viajantes individuais.



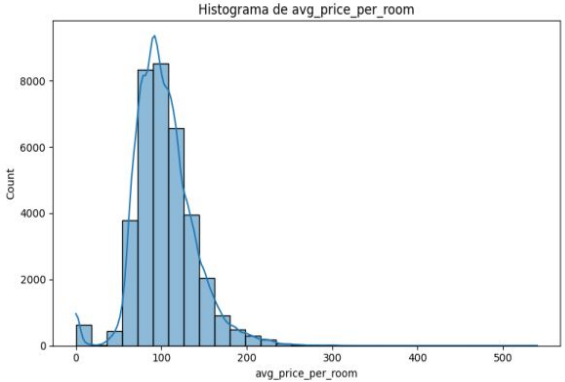
- Mostra como as reservas estão distribuídas em relação às noites de fim de semana.
- Pode indicar preferências por estadias durante a semana ou nos fins de semana.



- Analisa a distribuição das reservas em relação às noites de semana.
- Pode revelar se a maioria das reservas é para viagens de negócios ou lazer durante a semana.



- Exibe a distribuição do tempo entre a data da reserva e a data de chegada.
- Ajuda a entender padrões de reservas de última hora ou planejadas com antecedência.



- Mostra como os preços dos quartos estão distribuídos.
- Pode fornecer insights sobre a faixa de preço preferida pelos clientes.

Avaliação de modelos

Utilizamos algumas métricas para analisar qual o melhor modelo dentre os testados, que estão listadas no slide a seguir.

Métrica	O que Mede	Como é Calculada	Uso
Acurácia	Proporção de acertos do modelo.	(Verdadeiros Positivos + Verdadeiros Negativos) / Total de Previsões.	Avaliação geral de desempenho.
Pontuação F1	Equilíbrio entre precisão e recall.	2 * (Precisão * Recall) / (Precisão + Recall).	Útil em conjuntos de dados desbalanceados.
Precisão	Exatidão das previsões positivas.	Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Positivos).	Importante quando os falsos positivos são custosos.
Recall	Capacidade de identificar casos positivos.	Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Negativos).	Crucial quando os falsos negativos são custosos.

Resultados

Testamos 3 modelos diferentes, de 3 paradigmas diferentes, e avaliamos o desempenho de cada um, com o objetivo de averiguarmos qual modelo se adequa melhor aos dados apresentados.

1. Random Forest
2. KNN
3. Naive Bayes

Resultados:

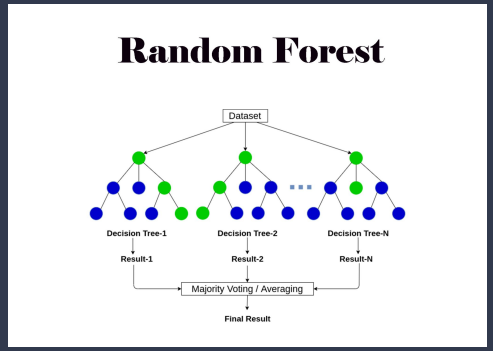
1) Random Forest

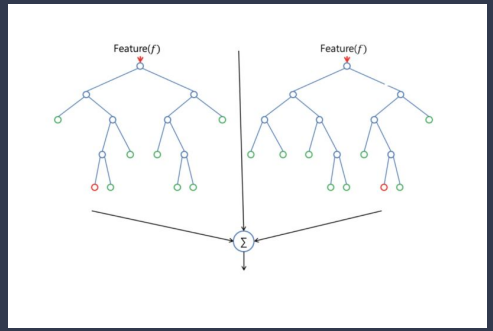
Paradigma: Aprendizado Simbólico + Ensemble

Ao combinar diversas árvores, cada uma treinada em subconjuntos diferentes dos dados, o modelo oferece estabilidade, reduzindo overfitting e proporcionando uma visão abrangente das relações nas variáveis.

Desempenho do Modelo:

	Model	Accuracy Score	F1 score	Precision	Recall
0	Random forest	0.904179	0.849263	0.893757	0.808989





Resultados:

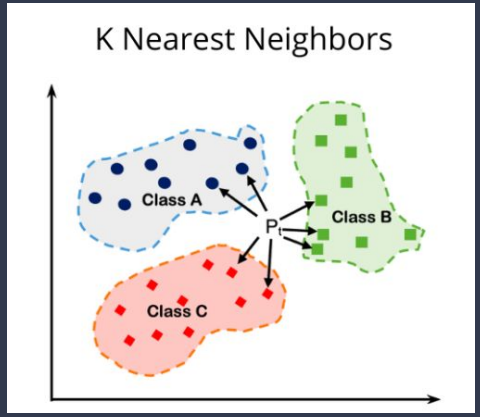
2) KNN

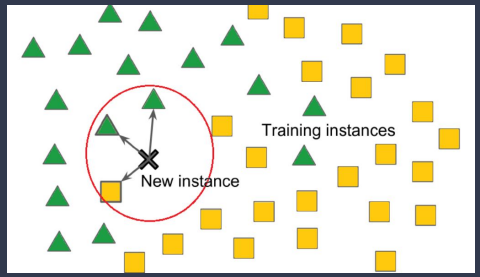
Paradigma: Aprendizado baseado em instâncias

O KNN, é uma técnica de aprendizado de máquina que se baseia na proximidade entre os pontos de dados. Ao classificar ou prever com base nos vizinhos mais próximos, o KNN destaca-se por sua simplicidade conceitual e eficácia em identificar padrões em conjuntos de dados.

Desempenho do Modelo:

	Model	Accuracy Score	F1 score	Precision	Recall
0	KNN	0.876172	0.808329	0.835863	0.782551





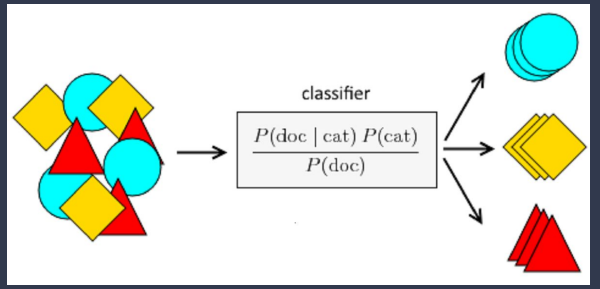
Resultados:

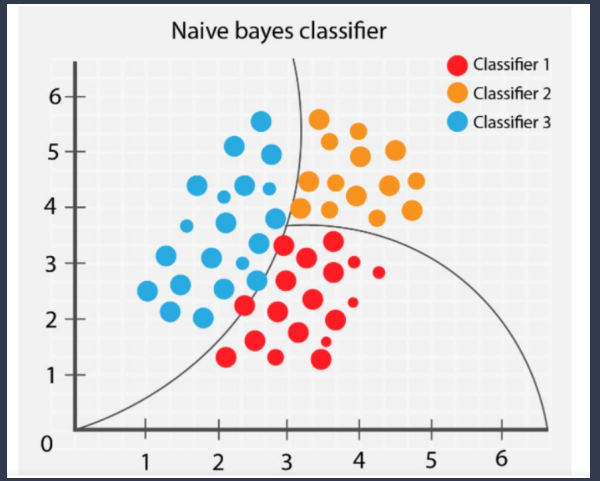
3) Naive Bayes: Categorical Naive Bayes

Paradigma: Aprendizado estatístico

Baseado no teorema de Bayes, o algoritmo calcula a probabilidade condicional de uma classe dado um conjunto de características. É especialmente útil quando as variáveis são independentes (hipótese ingênua).

O Naive Bayes categórico foi escolhido a partir da análise exploratória, onde foi percebida a grande quantidade de variáveis categóricas presentes nos dados.





Desempenho do Modelo:

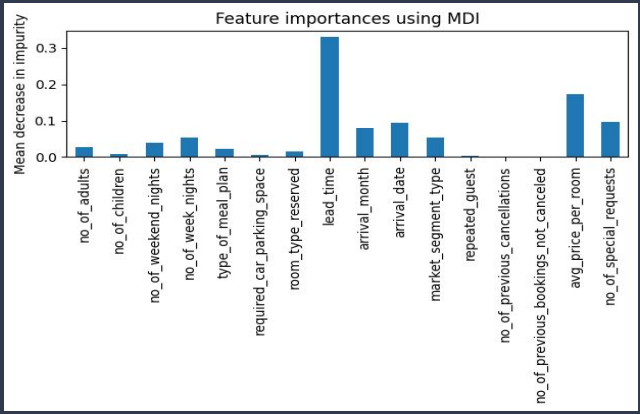
	Model	Accuracy Score	F1 score	Precision	Recall
0	Categorical_nb	0.801522	0.671293	0.729365	0.621786

Modelo escolhido:

Random Forest

Interpretabilidade das Features:

As barras no gráfico representam a importância de cada feature na previsão do modelo, calculada usando o critério Mean Decrease in Impurity (MDI).



Esta decisão foi motivada por sua boa performance, expressa por uma Accuracy Score de 90.4%, indicando sua eficácia em prever com precisão se um hóspede irá cancelar ou não.

O Random Forest também se destacou com um F1 Score de 84.9%, uma métrica essencial quando lidamos com desequilíbrio nos dados, como é comum em previsões de cancelamentos. Esta pontuação leva em consideração tanto a precisão quanto o recall, equilibrando a preocupação com falsos positivos e falsos negativos.

Além de sua performance sólida, o Random Forest é conhecido por lidar bem com diferentes tipos de dados, oferecendo robustez contra o sobreajuste e se mostrando eficaz em tarefas de classificação. Dessa forma, acreditamos que o Random Forest é a escolha mais adequada para atender às nossas necessidades específicas de previsão de cancelamentos de reservas.

Aplicações práticas dos resultados:

- Munida das análises aqui mostradas as empresas hoteleiras podem tomar diversas ações:
  - Dada a predição da reserva em tempo real, um prazo, valor e multas adaptativas podem ser implementadas;
  - Avaliando as estatísticas de cancelamento das reservas, novos métodos de contato com o cliente podem ser pensados, retomando a proximidade de antigamente.
  - Pode-se explorar também diferentes fontes de reserva, uma vez que uma empresa pode ter mais de um canal de atendimento.

Fonte dos dados

https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset/data

Integrantes:

Ana Vitória Gouvea de Oliveira Freitas

Eduardo Vinícius Barbosa Rossi

Matheus Luis Oliveira da Silva

Pedro Augusto Ribeiro Gomes

Sofhia de Souza Gonçalves

Thiago Henrique dos Santos Cardoso