UNIVERSIDADE DE SÃO PAULO
BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E COMPUTAÇÃO - ICMC

Gustavo Bartholomeu Trad Souza
Nº USP – 11219216

Desafio de Classificação

São Paulo 2023

Índice

Jupyter Notebook com código:
https://drive.google.com/file/d/1tZpMlG0-74gXQJJSrLS9IiR3Dj2zHRxy/view?usp=sharing

1. Descrição do problema e análise dos dados

Os canais online de reserva de hotéis mudaram drasticamente as possibilidades de reserva e o comportamento dos clientes em relação à reserva de vagas. Um número significativo de reservas em hotéis é cancelado e os motivos típicos para cancelamentos incluem mudança de planos, conflitos de agendamento, etc. Uma forma de evitar grandes despesas por parte dos hotéis é tentar prever a possibilidade do cancelamento de acordo com informações fornecidas durante a reserva, dessa forma, quando um cliente.

1.1. Descrição dos dados

O dataset é estruturado e é composto de 19 colunas com informações sobre reservas realizadas por clientes e algumas informações sobre o cliente que realizou a reserva. Não ha valores faltantes em nenhuma das colunas do dataset. As colunas são:

- **Booking_ID:** Identificador único de cada reserva.
- **no_of_adults:** Quantidade de adultos.
- **no_of_children:** Quantidade de crianças.
- **no_of_weekend_nights:** Quantidade de noites de fim de semana (sábado ou domingo) que o hóspede reservou.
- **no_of_week_nights:** Quantidade de noites de dias de semana (segunda a sexta) que o hóspede reservou.
- **type_of_meal_plan:** Tipo de plano de refeição reservado pelo cliente.
- **required_car_parking_space:** Necessita de vaga no estacionamento? (0-Não, 1-Sim).
- **room_type_reserved:** Tipo de quarto reservado pelo cliente. Os valores são cifrados (codificados) pela INN Hotels.
- **lead_time:** Quantidade de dias entre a data da reserva e a data de chegada.
- **arrival_year:** Ano da reserva.
- **arrival_month:** Mês da reserva.
- **arrival_date:** Dia do mês da reserva.
- **market_segment_type:** Designação do segmento de mercado.
- **repeated_guest:** O cliente ja se hospedou no hotel antes? (0-Não, 1-Sim).
- **no_of_previous_cancellations:** Número de reservas anteriores que foram canceladas pelo cliente antes da reserva atual.
- **no_of_previous_bookings_not_canceled:** Número de reservas anteriores que não foram canceladas pelo cliente antes da reserva atual.
- **avg_price_per_room:** Preço médio por dia da reserva; os preços dos quartos são dinâmicos. (em euros).
- **no_of_special_requests:** Número total de solicitações especiais feitas pelo cliente (por exemplo, número do andar, vista do quarto, etc.).
- **booking_status:** Flag indicando se a reserva foi cancelada ou não.

1.2. Atributos quantitativos e categóricos

O dataset possui 36275 entradas distintas e sem valores faltantes com informações de reservas realizadas por clientes em diferentes hotéis. Dos 19 atributos presentes na tabela, 11 são quantitativos, sendo eles: **no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, no_of_previous_cancellations, arrival_month, arrival_date, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests** e 6 atributos qualitativos, sendo eles: **type_of_meal_plan, required_car_parking_space, room_type_reserved, market_segment_type, repeated_guest e booking_status**. O atributo de identificação Booking_ID não é elevado em conta pois identifica univocadamente cada entrada da tabela e, portanto, não traz informações relevantes para a análise dos dados. O atributo arrival_year também será removido pois não traz informações que permitam prever o cancelamento de uma reserva uma vez que não se repete em dados futuros (de outros anos). O atributo binário booking_status é o atributo alvo que deve ser predito pelos algorpitmos de classificação que serão utilizados. Os atributos categóricos possuem as seguintes quantidades de valores únicos:

type_of_meal_plan: 4 - Meal Plan 1, Not Selected, Meal Plan 2, Meal Plan 3

room_type_reserved: 7 - Room_Type 1, Room_Type 4, Room_Type 2, Room_Type 6, Room_Type 5, Room_Type 7, Room_Type 3

market_segment_type: 5 - Offline, Online, Corporate, Aviation, Complementary

required_car_parking_space: 2 - 0, 1

repeated_guest: 2 - 0, 1

booking_status: 2 - Not_Canceled, Canceled

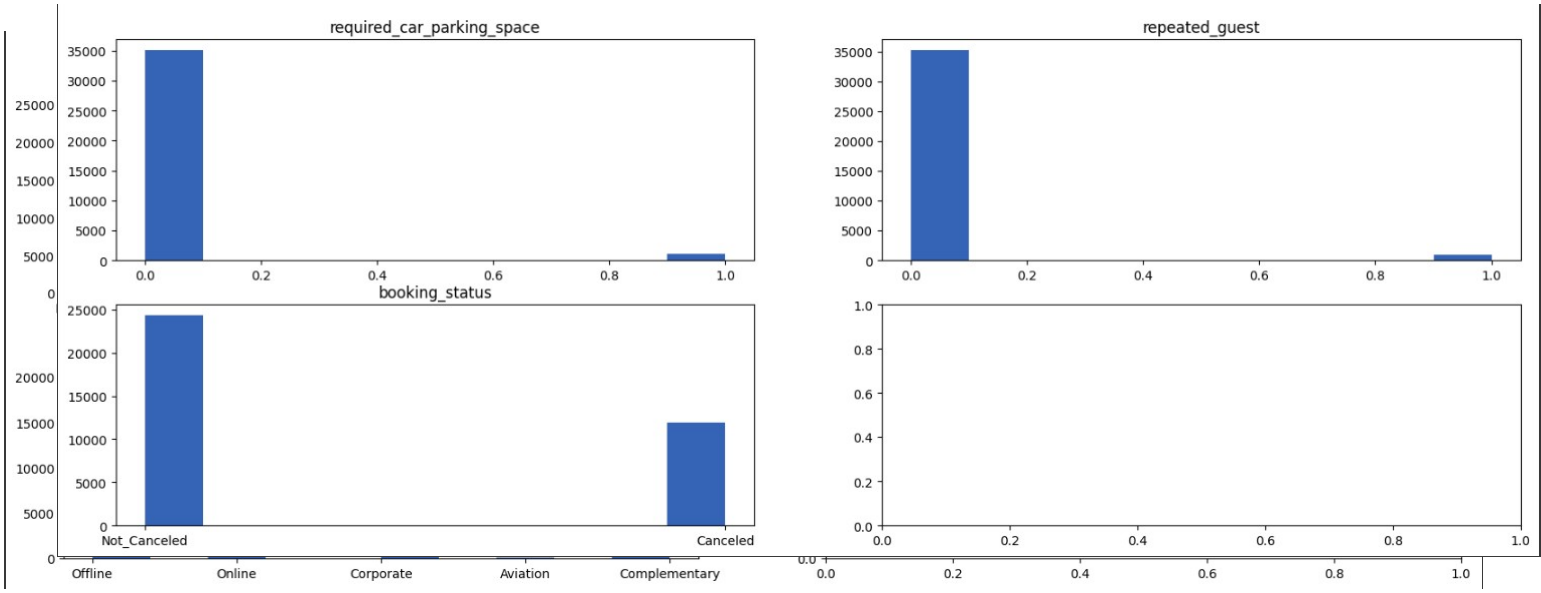
1.3. Correlação entre atributos quantitativos

Os atributos quantitativos possuem uma correlação muito baixa, sendo a maior correlação entre os atributos no_of_previous_cancellations e no_of_previous_bookings_not_canceled com o valor de 0.468147, portanto todos os valores serão mantidos. Abaixo segue a tabela de correlação dos atributos:

	required_car_parking_space	repeated_guest	Meal Plan 1	Meal Plan 2	Meal Plan 3	Not Selected	Room_Type 1	Room_Type 2	Room_Type 3	Room_Type 4	Room_Type 5	Room_Type 6	Room_Type 7	Aviation	Complementary	Corporate	Offline	Online
required_car_parking_space	1.00	0.11	0.03	-0.03	0.01	-0.00	-0.04	0.01	-0.00	0.02	-0.00	0.04	0.01	0.01	0.03	0.08	-0.10	0.05
repeated_guest	0.11	1.00	0.07	-0.05	0.01	-0.05	0.04	-0.01	-0.00	-0.04	0.02	-0.02	0.04	0.04	0.20	0.42	-0.07	-0.18
Meal Plan 1	0.03	0.07	1.00	-0.57	-0.02	-0.74	-0.22	0.06	-0.00	0.19	0.03	0.06	0.03	0.03	0.04	0.13	-0.01	-0.06
Meal Plan 2	-0.03	-0.05	-0.57	1.00	-0.00	-0.13	0.09	-0.03	-0.00	-0.07	-0.01	-0.01	-0.02	-0.02	-0.02	-0.07	0.30	-0.24
Meal Plan 3	0.01	0.01	-0.02	-0.00	1.00	-0.00	-0.02	-0.00	-0.00	0.00	-0.00	-0.00	0.11	-0.00	0.09	-0.00	-0.00	-0.02
Not Selected	-0.00	-0.05	-0.74	-0.13	-0.00	1.00	0.20	-0.04	0.01	-0.17	-0.03	-0.06	-0.03	-0.02	-0.04	-0.09	-0.23	0.27
Room_Type 1	-0.04	0.04	-0.22	0.09	-0.02	0.20	1.00	-0.26	-0.03	-0.83	-0.16	-0.31	-0.12	-0.04	-0.04	0.08	0.23	-0.24
Room_Type 2	0.01	-0.01	0.06	-0.03	-0.00	-0.04	-0.26	1.00	-0.00	-0.06	-0.01	-0.02	-0.01	-0.01	0.02	-0.03	-0.06	0.07
Room_Type 3	-0.00	-0.00	-0.00	-0.00	-0.00	0.01	-0.03	-0.00	1.00	-0.01	-0.00	-0.00	-0.00	-0.00	0.04	0.01	-0.00	-0.01
Room_Type 4	0.02	-0.04	0.19	-0.07	0.00	-0.17	-0.83	-0.06	-0.01	1.00	-0.04	-0.07	-0.03	0.06	-0.01	-0.08	-0.19	0.21
Room_Type 5	-0.00	0.02	0.03	-0.01	-0.00	-0.03	-0.16	-0.01	-0.00	-0.04	1.00	-0.01	-0.01	-0.01	0.04	0.08	0.00	-0.05
Room_Type 6	0.04	-0.02	0.06	-0.01	-0.00	-0.06	-0.31	-0.02	-0.00	-0.07	-0.01	1.00	-0.01	-0.01	0.01	-0.04	-0.10	0.11
Room_Type 7	0.01	0.04	0.03	-0.02	0.11	-0.03	-0.12	-0.01	-0.00	-0.03	-0.01	-0.01	1.00	-0.00	0.15	-0.01	-0.04	0.01
Aviation	0.01	0.04	0.03	-0.02	-0.00	-0.02	-0.04	-0.01	-0.00	0.06	-0.01	-0.01	-0.00	1.00	-0.01	-0.01	-0.04	-0.08
Complementary	0.03	0.20	0.04	-0.02	0.09	-0.04	-0.04	0.02	0.04	-0.01	0.04	0.01	0.15	-0.01	1.00	-0.03	-0.07	-0.14
Corporate	0.08	0.42	0.13	-0.07	-0.00	-0.09	0.08	-0.03	0.01	-0.08	0.08	-0.04	-0.01	-0.01	-0.03	1.00	-0.16	-0.32
Offline	-0.10	-0.07	-0.01	0.30	-0.00	-0.23	0.23	-0.06	-0.00	-0.19	0.00	-0.10	-0.04	-0.04	-0.07	-0.16	1.00	-0.85
Online	0.05	-0.18	-0.06	-0.24	-0.02	0.27	-0.24	0.07	-0.01	0.21	-0.05	0.11	0.01	-0.08	-0.14	-0.32	-0.85	1.00

1.4. Balanceamento dos dados

Os dados categóricos possuem algum desbalanceamento, com os atributos possuindo valores discrepantes entre a quantidade de elementos em cada classes. Os valores alvo possuem uma classe (Not_Canceled) com o dobro de elementos da outra (Canceled) portanto será utilizado o **SMOTE** para gerar mais exemplos da classe menos presente. Seguem os histogramas dos atributos categóricos:



1.5. Normalização dos atributos quantitaivos

Será aplicada a normalização min-max nas colunas quantitativas de forma que os valores ficarão entre 0 e 1.

1.6. Codificação dos atributos categóricos

Os dados categóricos assumem valores de strings que podem variar entre valores definidos. Essa forma de representação não é facilmente aplicável em algoritimos de classificação, onde dados numéricos são preferíveis, portanto uma forma de solucionar esta questão é transformar cada categoria em uma coluna e o seu valor assume 0 nas entradas que não pertencem àquela classe e 1 para as que pertencem. Dessa forma eles assumem valores numéricos que mantêm a semantica alterando apenas a forma de representação dessas informações.

2. Descrição das técnicas utilizadas

A tarefa pode ser abordada como um problema de classificação, onde a partir das informações de uma determinada reserva um algoritimo é treinado para definir se aquela reserva

será ou não cancelada. A classe de saída é interpretada como tendo o valor 0 para as reservas não canceladas e 1 para as reservas canceladas. Os dados possuem tanto atributos quantitativos quanto categóricos (transformados em classes binárias). Apesar de ter sido realizado um balanceamento na quantidade dos atributos alvo, os demais atributos categóricos possuem uma certa quantidade de desbalanceamento que podem interferir negativamente nos algoritimos selecionados. Os algoritimos foram treinados e avaliados no dataset desbalanceado e balanceado a fim de observar a alteração no desempenho causado pelo balanceamoto do dataset em relação ao atributo alvo.

2.1. KNN

O algorítimo de K-Nearest Kneighbour foi escolhido pois é uma técnica simples porém eficaz de classificação que classifica os dados em conjuntos baseados na distância entre os atributos. O número de conjuntos k nos quais os dados serão separados é um parâmetro importante pois define a qualidade do agrupamento que será realizado. Serão testados diferentes valors de k, sendo eles 1, 2, 3, 11 e 20.

2.2. Support Vector Machine (SVM) com Polinômio de grau N

Essa técnica consiste em separar os pontos que pertencem à classes diferentes encontrando os hiperparâmetros que melhor descrevem um hiperplano de grau N separando os dados. A separação pode ser linear ou não linear. Devido à quantidade de atributos sendo avaliados à fim de determinar se uma reserva será cancelada, foi escolhida a máquina que descreve separações não lineares, dando um poder maior de representação desse domínio mais complexo da aplicação, que talvez fosse prejudicado por uma separação linear. Apesar disso o grau 1 foi utilizado testando também a possibilidade linear. Os graus sendo testados serão 1, 2 e 5.

2.3. Random Forest

É um método de Bagging baseado em árvores de decisão que se utiliza de N árvores em paralelo treinadas com sunconjuntos do conjunto de treino, facilitando a identificaçõa de relacionamentos entre as variáveis. Os números de árvores na forest para cada teste foi definido como 10, 20,50 e 100, o critério de avaliaçõ do split foi o Gini para todos os testes e a profundidade máxima foi definida como sendo metade do número de atributos dos dados de entrada.

2.4. Multi-layer Perceptron (MLP)

É uma técnica que consiste em acplar sequencialmente conjuntos de Perceptorns a fim de criar um aproximadpr de funções não linear. Através de operações com paramêtros ajustáveis chamados pesos e bias os dados de entrada são transformados em uma saída que é uma predição do atributo alvo, em seguida o erro entre o valor real e apredição é calculado e à partir desse erro é calculado o gradiente negativo (pois o objetivo é reduzir o erro) para cada parãmetro ajustável e os ajustes são aplicados. O modelo escolhido é um modelo simples com 3 camadas, sendo a primeira e a ultima as camas de entrada e saída respectivamente, portanto o número de elementos em cada uma é o número de atributos de entrada ou o número de atributos sendo classificados e número de elementos na camada oculta de cada modelo de teste foi respectivamente 10, 20, 50 e 100.

3. Interpretação dos resultados obtidos

3.1. KNN

Dados desbalanceados - Acurácia com 1 K-NN: 0.8495 +/- 0.0052
Dados balanceados - Acurácia com 1 K-NN: 0.9937 +/- 0.0013

Dados desbalanceados - Acurácia com 2 K-NN: 0.8501 +/- 0.0031
Dados balanceados - Acurácia com 2 K-NN: 0.9903 +/- 0.0012

Dados desbalanceados - Acurácia com 3 K-NN: 0.8497 +/- 0.0028
Dados balanceados - Acurácia com 3 K-NN: 0.9849 +/- 0.0017

Dados desbalanceados - Acurácia com 11 K-NN: 0.8422 +/- 0.0040
Dados balanceados - Acurácia com 11 K-NN: 0.9454 +/- 0.0030

Dados desbalanceados - Acurácia com 20 K-NN: 0.8388 +/- 0.0033
Dados balanceados - Acurácia com 20 K-NN: 0.9238 +/- 0.0041

3.2. SVM

Dados desbalanceados - Acurácia com SVM Poly grau 1 : 0.8005 +/- 0.0024
Dados balanceados - Acurácia com SVM Poly grau 1 : 0.8415 +/- 0.0048

Dados desbalanceados - Acurácia com SVM Poly grau 2 : 0.8122 +/- 0.0042
Dados balanceados - Acurácia com SVM Poly grau 2 : 0.8687 +/- 0.0027

Dados desbalanceados - Acurácia com SVM Poly grau 5 : 0.8273 +/- 0.0047
Dados balanceados - Acurácia com SWVM Poly grau 5 : 0.8929 +/- 0.0036

3.3. Random Forest

Dados desbalanceados - Acurácia Random Forest 10 estimadores: 0.8766 +/- 0.0033
Dados balanceados - Acurácia Random Forest 10 estimadores: 0.9504 +/- 0.0028

Dados desbalanceados - Acurácia Random Forest 20 estimadores: 0.8799 +/- 0.0021
Dados balanceados - Acurácia Random Forest 20 estimadores: 0.9546 +/- 0.0024

Dados desbalanceados - Acurácia Random Forest 50 estimadores: 0.8822 +/- 0.0018
Dados balanceados - Acurácia Random Forest 50 estimadores: 0.9545 +/- 0.0028

Dados desbalanceados - Acurácia Random Forest 100 estimadores: 0.8822 +/- 0.0037
Dados balanceados - Acurácia Random Forest 100 estimadores: 0.9543 +/- 0.0032

3.4. MLP

Dados desbalanceados - Acurácia MLP 10 hidden units: 0.8202 +/- 0.0072
Dados balanceados - Acurácia MLP 10 hidden units: 0.8761 +/- 0.0069

Dados desbalanceados - Acurácia MLP 20 hidden units: 0.8259 +/- 0.0047
Dados balanceados - Acurácia MLP 20 hidden units: 0.8888 +/- 0.0050

Dados desbalanceados - Acurácia MLP 50 hidden units: 0.8401 +/- 0.0047
Dados balanceados - Acurácia MLP 50 hidden units: 0.9118 +/- 0.0047

Dados desbalanceados - Acurácia MLP 100 hidden units: 0.8467 +/- 0.0059
Dados balanceados - Acurácia MLP 100 hidden units: 0.9258 +/- 0.0062

Dentre os métodos utlizados para analisar os dados de reservas de hospedagem e prever a possibilidade de cancelamento da reserva, o que obteve a maior acurácia calculada por cross_validation foi o KNN com número de vizinhos igual a 1, apresentando 99% de acurácia seguido pelo KNN com k = 2. A técnica de Random Forest obteve 95 % de acurácia e o MLP obteve, na sua melhor versão, 92% de acurácia e por fim a técnica de SVM obteve acuracias inferiores à 80%.

O método de KNN pode ter sido mais efetivo pois é menos afetado pela grande quantidade de atributos e pelo desbalanceamento de alguns atributos categóricos, uma vez que por possuiem apenas os valores 0 ou 1 a distância naquela dimensão entre dados diferentes é significativa e a fronteira pode ser facilmente estabelecida. Os métodos de Random Forest e MLP alcançaram resultados próximos, o método de Random Forest atingiu valores próximos de acurácia com 20, 50 e 100 árvores, portanto a partir de determinado número de árvores, a complexidade computacional não compensa o melhor desempenho do modelo. O método de MLP obteve uma acurácia de 92% na sua melhor configuração com 100 elementos na camada oculta porém o tempo de treinamento superou muito os demais métodos, portanto para datasets mais simples com muitos elementos categóericos ele pode não ser ideal. O método de SVM pode ter obtido os piores resultados por conta do grau baixo da função responsável por gerar o hiperplano, com o aumento do grau é possível observar que a acuracia aumentou porém o custo necessário também aumentam deixando o algoritimo significativamente mais lento que os demais.

4. Conclusão

Os resultados obtidos experimentando diferentes configurações dos métodos de KNN, SVM Random Forest e MLPs permitem concluir que apesar de ser mais símples o método do KNN se mostrou mais efetivo, mostrando que a complexidade do método não está diretamente relacionado à sua eficácia. A diferença de tempo de alguns dos testes com SVMs e MLPs para os testes de KNN foi muito discrepante podendo atingir vários minutos para um pior desempenho. Outra observação importante é o efeito do balanceamento e pré processamento dos dados, que melhoraram significativamente a acurácia de todos os métodos, provavelmente devido ao overfitting que pode

ocorrer com o favorecimento da predição em direção à classe dominante e da dominação das classes com maior magnitude no caso da normalização.

5. Apêndice (descrição das técnicas de Classificação apresentadas pelos colegas durante os seminários)

5.1. Novo método de detecção de água envenenada usando tecnologia Wi-Fi incorporada a smartphone e algoritimos de aprendizado de máquina.

- Amostras de água foram submetidas a um sinal Wi-Fi e o comportamento do sinal foi observado a fim de medir, através das características do sinal recebido, se a água está contaminada ou não.
- **Pipeline do experimento:** Roteador -> Transmite sinal Wi-Fi através da amostra de água -> Smartphone recebe sinal -> Remoção de ruído do sinal -> Extração de CSI (Channel-State-Information) do sinal limpo de ruído -> Seleção de features a serem analisadas -> Alimentação do algoritimo de classificação -> Classificação de amostras entre contaminada ou não contaminada.
- **Algoritimos utilizados:**
  - SVM - kernel Linear e Gausiano apresentaram os melhores resutados.
  - KNN - k=1 foi a melhor opção.
  - LSTM - Dados foram normalizados através do z-score, duas camadas ocultas com 200 e 100 nós, dropout e softmax nas camadas.
  - Ensembles - utilizado o AdaBoost realizado de forma sequencial e com 30 iterações.
- **Métricas de avaliação:**
  - AUC - Área abaixo da curva ROC
  - TPR - taxa de posotivos verdadeiros
  - TNR - taxa de negativo verdadeiro
  - F1-Score - Combinação de precisão e TPR
  - Acurácia - Quantidade de acertos sobre o número total de predições
- **Resultados:**
  - água contaminada com 100mg/l e água limpa - pior método foi o KNN e o melhor foi o LSTM.
  - água contaminada com 1000mg/l e água limpa - taxa de acerto de 100% para todos os algoritimos.
  - água conatminada com 100mg/l, água contaminada com 1000mg/l e água limpa - AdaBoost obteve a maior acurácia e KNN obteve a pior.
- **Conclusão:**
  - Houve pouca diversidade em relação à toxina e às quantidades diluídas.

- Os hiperparâmetros das redes não foram divulgados.
- O experimento onde todos os modelos obtiveram 100% de acurácia deve ser mais analisado pois esse resultado distoa da prática encontrada na área.

5.2. Predição da Recidiva de Câncer usando técnicas de aprendizado de máquina.

- **Objetivo:** Realizar a predição da possibilidade de Recidiva de Cãncer à partir de features binárias relacionadas principalmente à presença de determinados tratamentos.
- Dataset muito desbalanceado - Apenas 9% das instância possuiam o rótulo de recidiva positiva.
- Para lidar com o desbalancemento foi utilizado K-Fold durante treino/teste, a fim de remover viés de sorte e SMOTE para realizar a geração de dados sintéticos para a classe com menos rótulos a fim de balancear os dados e evitar overfitting.
- **Modelos utilizados:**
  - Naive Bayes com BernoulliNB pois os atributos são binários e a distribuição de Bernoulli é utilizada para modelar problemas em que a resposta é binária.
  - SVM para encontrar fronteiras de decisão que melhor sepra as classes.
- **Resultados:**
  - O modelo Naive Bayes sofreu overfitting com os dados desbalanceados e houve um decrescimento das métricas ao ser ajustado com os dados balanceados.
  - O SVM apresentou um resultado melhor que o método de Naive Bayes porém também houve uma queda nas métricas com os dados balanceados. Além disso esse modelo possui pouca explicabilidade em relação ao resultado da classificação.
- **Conclusão:**
  - Trabalhos da área de saúde geralmente possuem um alto desbalanceamento, uma vez que medições saudáveis são mais abundantes que medições com não saudáveis.
  - O uso da sobreamostragem se provou bastante vantajoso nesse caso pois previniu o overfitting do modelo.

5.3. CatBoost

- A maioria dos algoritmos de ML trabalha melhor com atributos numéricos, portanto, a solução solução mais comum é transformar cada uma das colunas categóricas em uma colunas binárias para cada classe daquela categoria, indicando o pertencimento ou não da entrada àquela classe, porém para um número muito extenso de dados essa abordagem pode aumentar muito a dimensão e a disperssão dos dados.
- **CatBoosting (Categorical Boosting)** é uma técnica que lida com atributos categóricos sem pré-processamento. Ensemble do tipo Boosting onde os modelos utilizados são árvores de decisão.

- **Modelos utilizados para comparação:**
  - KNN
  - Random Forest
  - Naive Bayes
  - SVM
  - CatBoost
- **Vantagens:**
  - Maior desempenho com dados desbalanceados ou com alta dimensionalidade.
  - Lida bem com dados categóricos
  - Maior velocidade de treinamento
- **Desvantagens:**
  - Baixa interpretabilidade - Black Box
  - Seleção de hiperparametros pode ser complexa