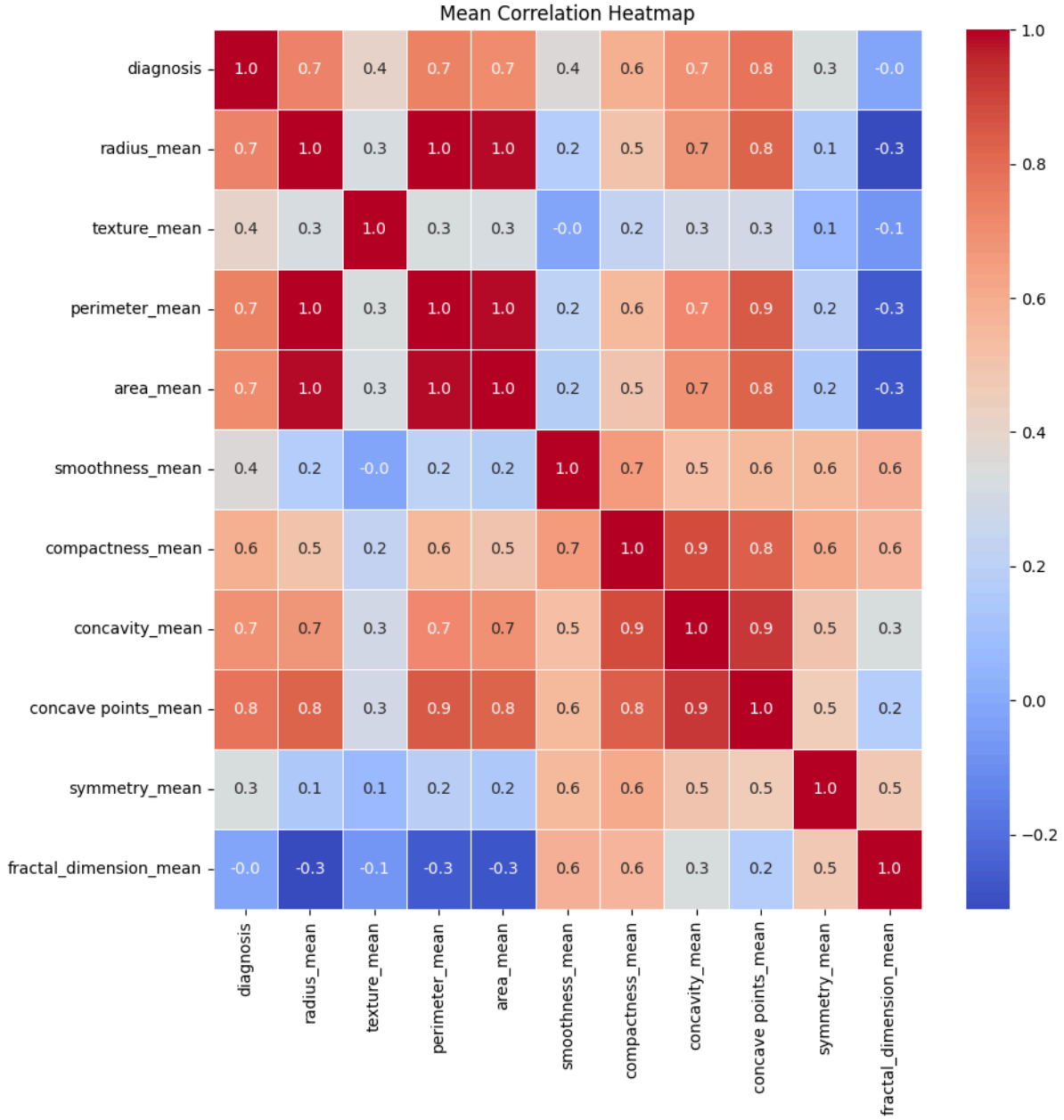
Tech challenge Fase 1

# Relatório Técnico

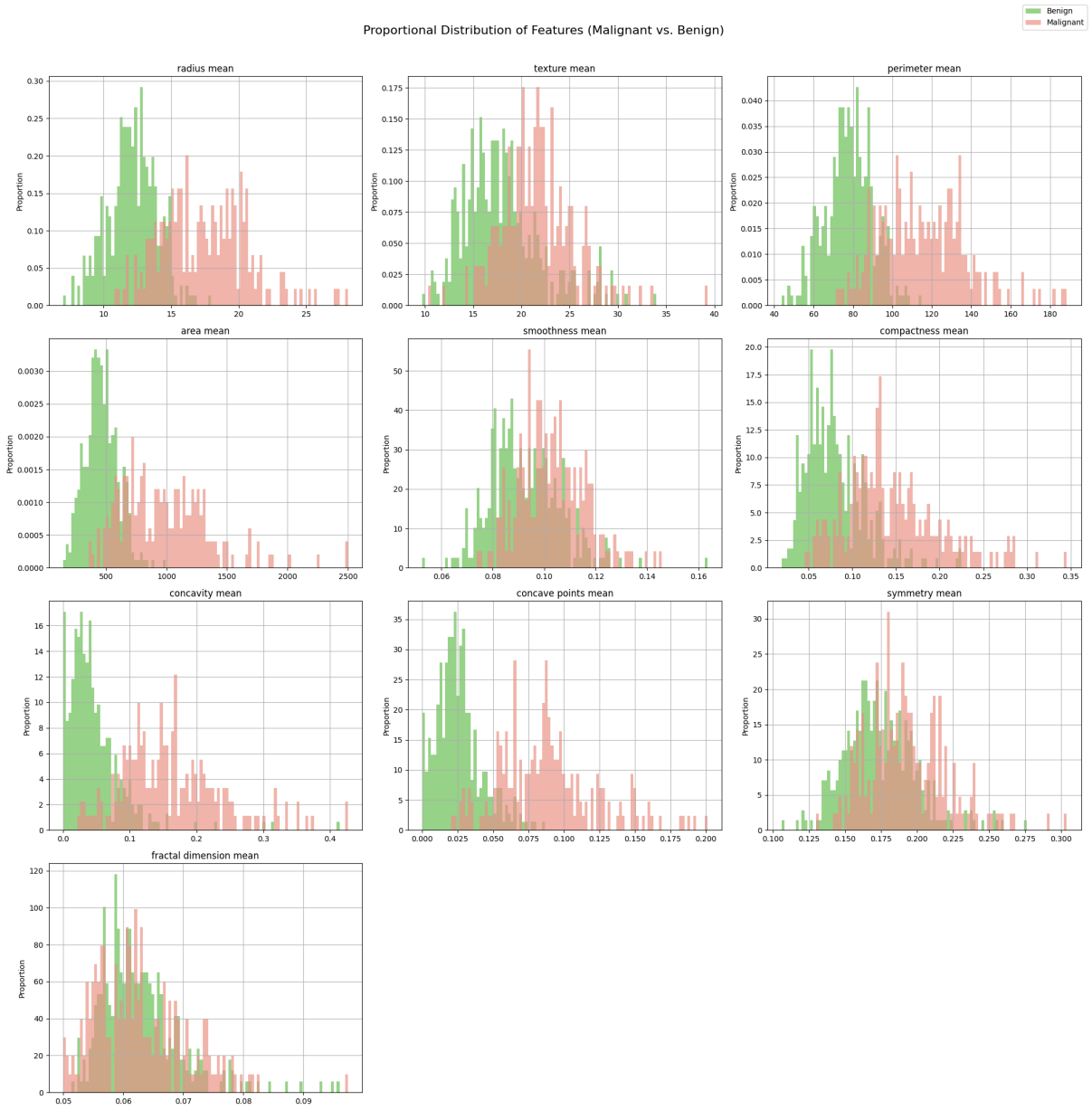
Foi utilizado um dataset simples que foi sugerido no próprio documento do Tech Challenge: Breast Cancer Wisconsin (Diagnostic) Data Set. É um dataset pequeno e com valores já tratados. Foi necessário apenas retirar duas colunas do dataset, pois uma é um id do paciente e outra uma coluna com todos os valores nulos.

Inicialmente, analisando a correlação entre os campos, fica clara a relação direta entre os campos que armazenam características da forma dos núcleos das células (raio, perímetro e área) e o diagnóstico, chegando a valores muito próximos a 1 quando exibida a correlação.




Inclusive há uma forte correlação entre todos os campos de área e também os campos que armazenam a forma do núcleo (concavidade, compactação e pontos côncavos). Fiz

algumas tentativas para tentar usar apenas um dos campos de cada um dos grupos para evitar multicolinearidade (por exemplo, retirar área, raio, concavidade e compactação) mas o resultado do treinamento dos modelos ficou pior.




Aqui há alguns gráficos que demonstram o quanto as colunas citadas acima influenciam no diagnóstico. Tanto que o treinamento do modelo usando XGBoost e RandomForest deram feature importance alta para estes campos e mais baixa para dimensão fractal, por exemplo.

A seleção dos modelos foi estratégica para cobrir as principais famílias de algoritmos de aprendizado supervisionado. O objetivo foi avaliar desde modelos clássicos e mais simples até arquiteturas mais complexas, como redes neurais, para este problema de classificação tabular.

- KNN: Como o modelo prevê seus targets por proximidade e o dataset é composto por valores muito bem definidos (principalmente em se tratando dos campos de área e formato da célula), esse pareceu um problema em que ele se sairia muito bem. A

ideia foi usar o KNN como nota de corte e que os outros modelos, principalmente usando redes neurais se sairiam melhor
- RandomForest e XGBoost: Me pareceram degraus acima do KNN, lidando com dados tabulares e para evitar overfitting sem necessariamente ter uma performance muito pior que o KNN, como redes neurais tem
- MLPClassifier: Queria testar como a rede neural se sairia comparando um paradigma tão diferente dos outros algoritmos, esperando inclusive que fosse se sair muito melhor do que todos os anteriores.

Dentre os resultados, o KNN e o MLPClassifier saíram-se melhor. Com vantagem neste caso para o KNN por ser um algoritmo muito mais rápido e com menos uso de recursos computacionais. Mas é provável que esta performance do KNN tenha sido tão surpreendente porque o dataset é muito pequeno (apenas 569 registros). Talvez com mais registros e mais variabilidade teríamos uma performance superior dos demais modelos.

Cabe destacar que o modelo com a maior precisão dos quatro foi o RandomForest, porém ele apresentou o pior recall, o que é péssimo para predições onde o prejuízo de um falso negativo é muito grande. Precisamente o caso de um diagnóstico médico.