UNIVERSIDADE FEDERAL DO RIO GRANDE – FURG

(Engenharia de Computação)

Relatório de Análise de Dados e Machine Learning

Pedro Henrique Fernandes

Matrícula: 124849

Rio Grande
2023

1. Introdução

O diagnóstico precoce e preciso do câncer é fundamental para melhorar os resultados do tratamento e aumentar as chances de sobrevivência do paciente. Contudo, muitas vezes a busca por um diagnóstico eficaz envolve uma série de exames invasivos e demorados, causando ansiedade e estresse para os pacientes e os profissionais da saúde.

Ao longo deste relatório, serão apresentadas técnicas de análise de dados e machine learn aprendidas em um bimestre de aulas de Sistemas Inteligentes utilizando o micro framework sklearn em python com o um objetivo claro: Reduzir o número de exames para o diagnóstico de um tipo específico de câncer. Para isso, vamos utilizar um data base adaptado que possui dados anotados de exames de câncer, disponibilizado em:

https://github.com/alura-cursos/reducao-dimensionalidade/blob/master/data-set/exames.csv

2. Descrição do Database

Antes de introduzir a metodologia, neste tópico será adotado uma prévia análise da conjunção do dataset.

O dataset consistem num arquivo CSV contendo cinco diagnósticos de câncer para 35 exames cada, assim compondo de 150 exames na base de dados. Portanto, temos como variável categórica o diagnóstico de câncer e variáveis numéricas as 35 baterias de exames.

3. Metodologia e Resultados Obtidos

Antes de introduzir a metodologia, neste tópico será adotado uma prévia análise da conjunção do database.

3.1 Workflow

Para a organização do documento

O workflow adotado possui quatro estados: Entrada de dados, pré-processamento e transformação nos dados, treinamento e construção do modelo ML e dados de teste.

Imagem 1: Estados do workflow adotado

3.2 Entrada de Dados

O primeiro estágio do workflow é a entrada de dados, que envolve a coleta ou obtenção dos dados anotados vindo do dataset de exames.

Após a entrada dos dados, foram empregadas as técnicas de remoção de celular considerando as análises de células vazias, features constantes, remoção de variáveis altamente correlacionadas e seleção automática de features utilizando a técnica *k-best*.

Essa etapa visa preparar os dados para serem utilizados pelo modelo de ML. Ela pode envolver a limpeza dos dados, como remoção de outliers e tratamento de valores discrepantes, normalização ou padronização das variáveis, além de outras técnicas específicas de acordo com o tipo de dado e o objetivo do modelo. O pré-processamento adequado é fundamental para garantir a confiabilidade e a eficácia do modelo.

3.2 Determinação do Baseline

Com a nossa base de dados setada, foi necessário dividir o conjunto de dados em dados de teste e dados de treinamento e determinar o *baseline*, ou seja, a acurácia da primeira versão do database e posteriormente aplicar as devidas reduções dimensionais. Para isso foi utilizado o método de classificação de floresta aleatória com 100 árvores classificadoras. Nesse modelo, cada árvore de decisão é criada utilizando porções geradas dos dados originais (treinamento) aleatoriamente. Cada árvore gera sua própria previsão é votada em um resultado. Assim sendo, o modelo de floresta considera votos de todas as árvores de decisão para prever ou classificar os resultados de uma amostra desconhecida, tal como esquematizado na figura abaixo:

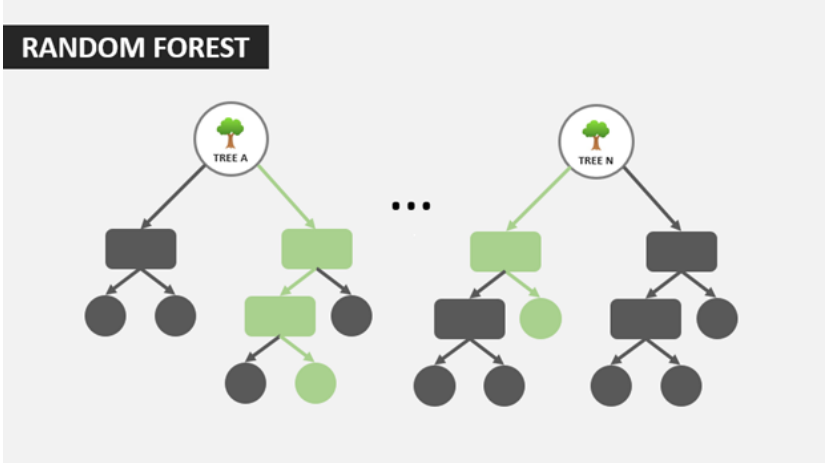

Imagem 2: Esquema do método de classificação random forest

Dessa forma foi obtido a acurácia de 98%. Afins de viés confirmatório da acurácia do classificador de floresta, foi implementado outro classificador que utiliza a heurística do mais frequente, em outras palavras para todos os dados o

classificador considera o valor da variável categórica o mais frequente. Com isso obtemos uma acurácia de 66,66%. Dessa forma confirmando que temos um ótimo baseline.

3.3 Análise de Dados e Remoção Dimensional Manual

3.3.1 Dropando Dados Constantes

Foi plotado um gráfico violino utilizando os dados anotados e o resultado obtido de uma amostra dos dados pode ser analisado abaixo:

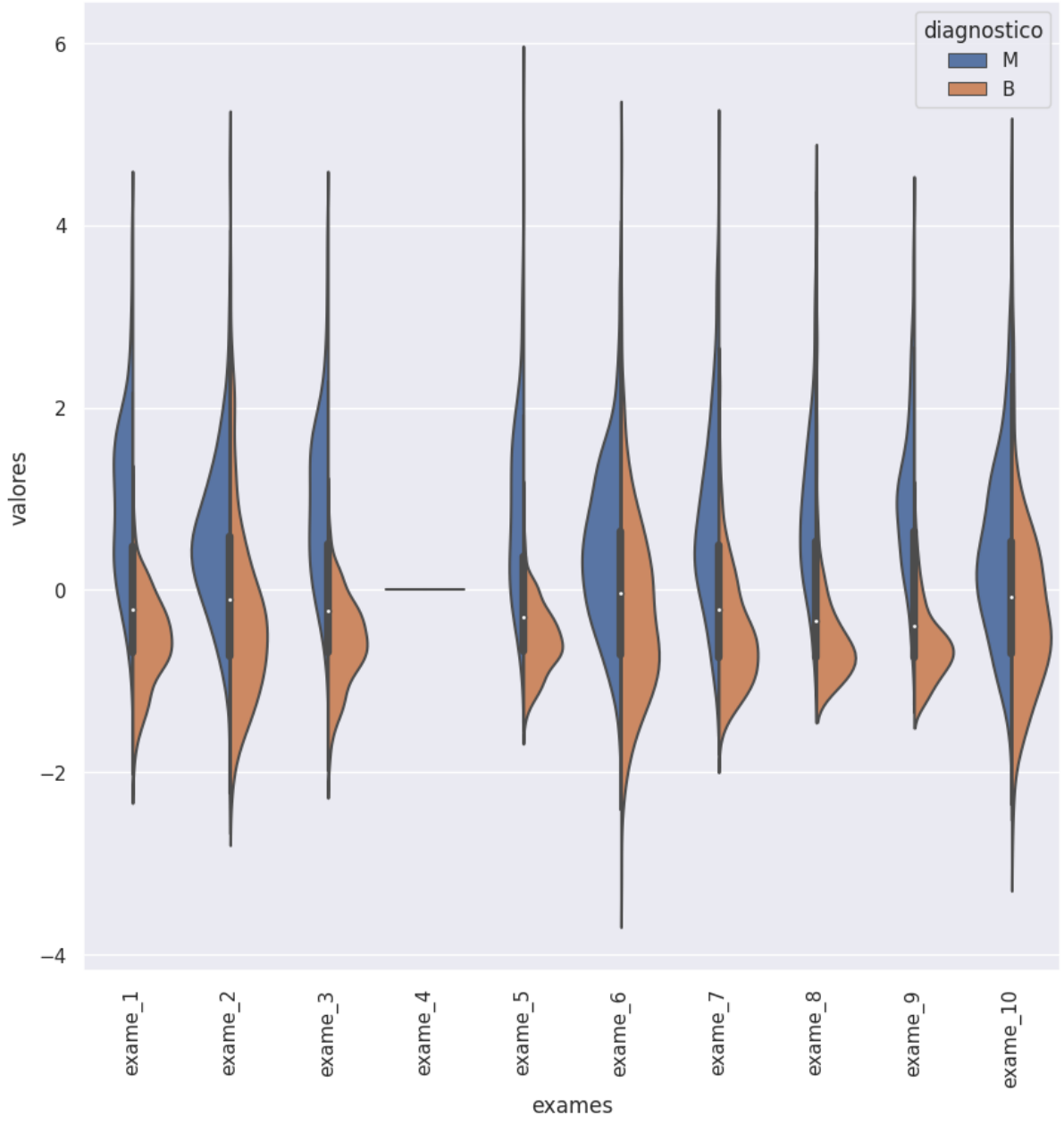

Imagem 4: Gráfico violino do dataset de câncer com valor constante no exame 4

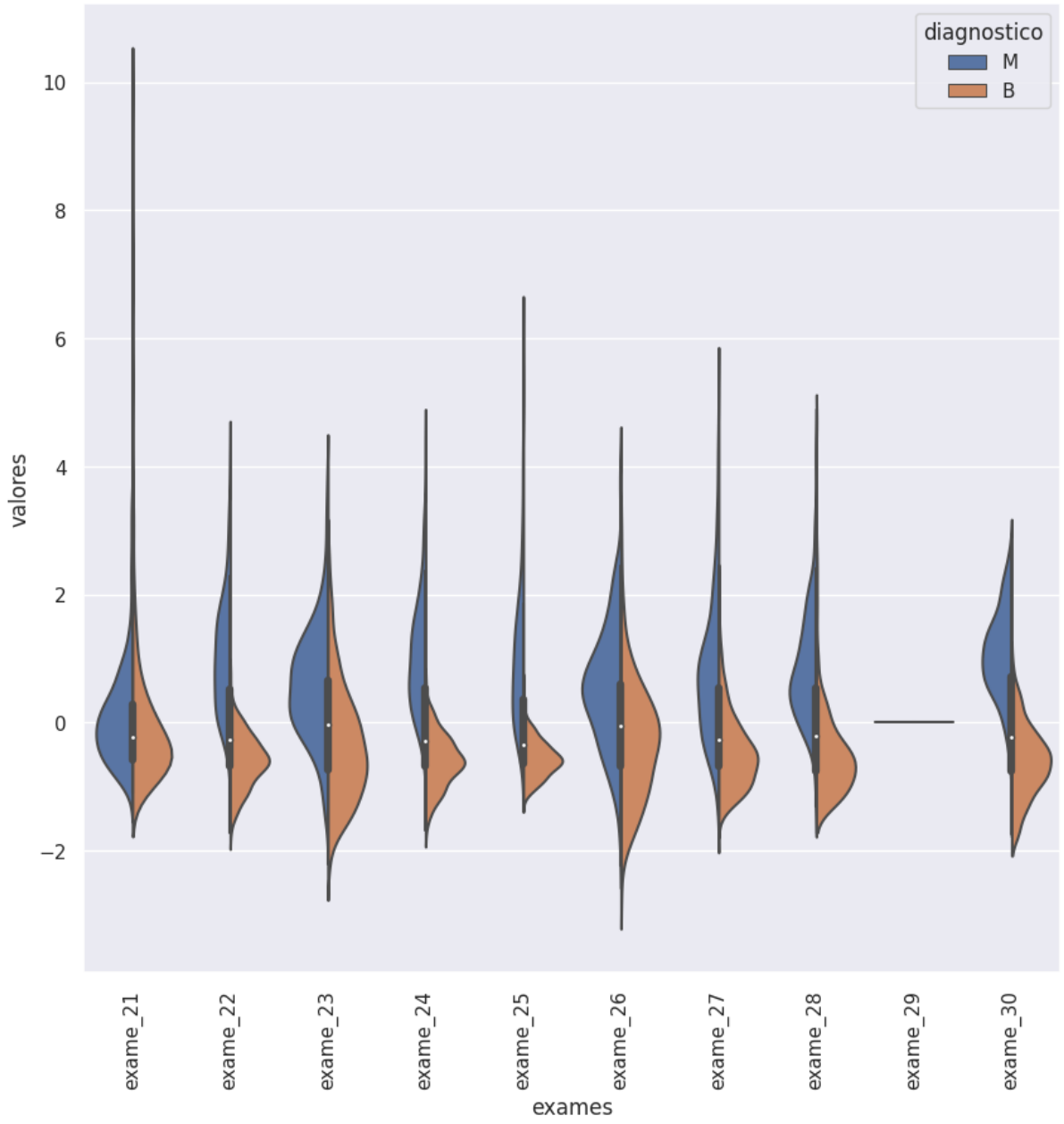

Imagem 5: Gráfico violino do dataset de câncer com valor constante no exame 29

Como podemos notar, obtivemos valores constantes em dois exames, no exame 4 e no exame 29, respectivamente. Dessa forma, possibilitando a remoção de ambas as features do nosso conjunto de dados anotados.

Calculando a nova acurácia do dataset com essas features removidas obtemos a seguinte acurácia: *Accuracy = 95%*

3.3.2 Dropando Dados Altamente Correlacionados

Para reduzir mais dimensões, foi realizada análise no mapa de calor com o objetivo de analisar a possibilidade de variáveis altamente correlacionadas.

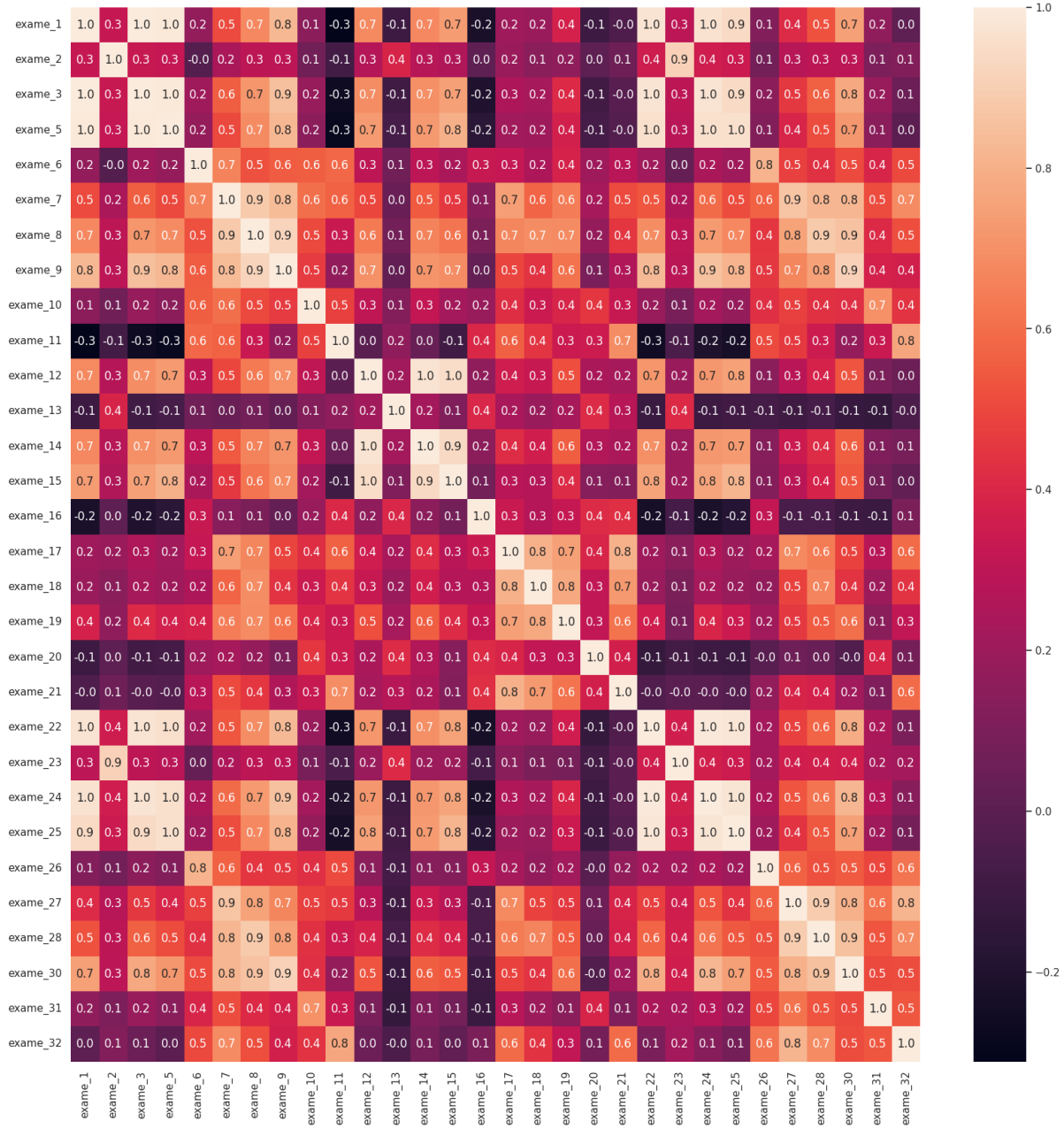

Imagem 6: Gráfico de calor de correlação de variáveis

A Fim de facilitar a análise do gráfico, foi impresso na tela os valores de variáveis com correlação maior que 1, como é mostrado no código abaixo:

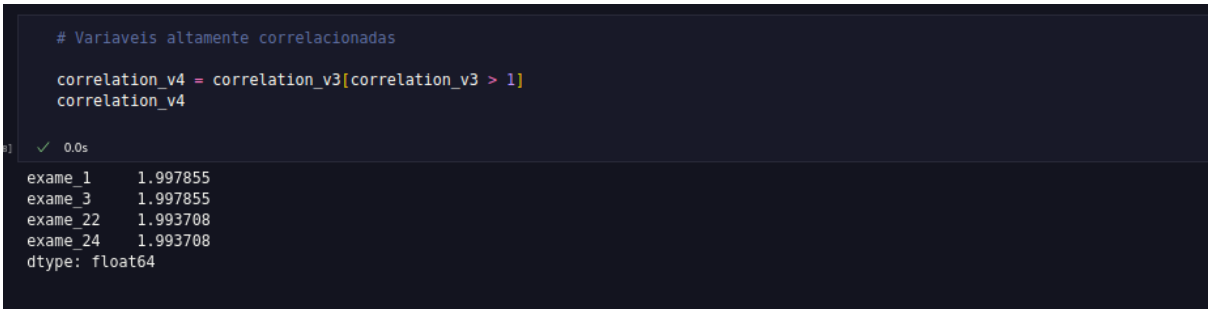

Imagem 7: Exames altamente correlacionados

Obtemos 4 exames altamente correlacionados, dessa forma removemos dois e obtemos uma nova acurácia: *Accuracy 95%*

Enfim, de uma forma manual, conseguimos remover cinco exames para determinação de câncer. Infelizmente faltando 30 baterias de exames.

3.4 Análise de Dados e Remoção Dimensional Manual

A fim de reduzir de forma significativa o número de bateria de exames, foi implementado uma outra técnica de redução dimensional, chamada de *select-kbest*. Uma função fornecida pelo framework *sklearn* que seleciona as K melhores features, impactando minimamente na acurácia do classificador. Dessa forma, obtemos uma nova acurácia: *Accuracy 93%*

Plotando a matriz de confusão notamos que 108 cancers benignos, pré dizemos 105 corretamente. E de 63 cânceres malignos, pré dizemos 55 corretamente.

Referências

Porque eliminar todas as correlacionadas, e não metade?

https://cursos.alura.com.br/forum/topico-porque-eliminar-todas-as-correlacionadas-e-nao-metade-194104.

Curso de ML. https://cursos.alura.com.br/course/reducao-dimensionalidade.