Para a execução da tarefa primeiramente foi analisado que a base tem bastante dados faltantes, logo não seria uma boa ideia exclui-los. Para não perder as informações dos dados, foi procurado um método de realizar a imputação dos dados faltantes. Como os dados são quantitativos, um método adequado seria o KnnImpute, que será aplicado com o argumento k = 5. Esse método consiste em observar os k vizinhos de uma observação faltante e fazer uma média, esse valor será atribuído para essa observação.

A base principal foi dividia em amostra treino com 80% dos dados, e teste com os 20% restantes. A base treino foi treinada com os seguintes procedimentos: KnnImpute, citado anteriormente; Nzv (Near Zero Variation) que remove as variáveis com variância próximas ou iguais a 0; Coor que remove variáveis correlacionadas, foi utilizado os argumentos FreqCute = 18, UniqueCut = 10 e cutoff = 0.8. Como a base de dados não é considerado uma base com uma grande quantidade de variáveis, foi escolhido não aplicar o método de Análise de Componentes principais.

Para a escolha de melhor modelo foi criado o critério de escolha que leva em conta a Acurácia, Sensibilidade e Especificidade, da forma:

Para o treinamento de Árvores Aleatórias foram escolhidos os critérios mtry = 2, 3, 5, 8, 13 e ntree = 500, 1000, 1500. Para o procedimento de reamostragem foi utilizado o padrão OOB. Foram gerados os seguintes resultados para a amostra teste:



Para o treinamento de AdaBoost foram escolhidos os critérios maxdeepth = 1, 2, 3, 5, 8 e mfinal = 150, 300, 450. Não foi utilizado procedimento de reamostragem pois o tempo de processamento do treinamento do modelo aumentava bastante. Foram gerados os seguintes resultados para a amostra teste:



O modelo que apresentou a melhor métrica para o a coluna “Critério” foi o de AdaBoost com maxdeepth = 2 e mfinal = 450. Foi aplicado os dados de validação para esse modelo e foi obtido os seguintes resultados:



Foi observado melhores estatísticas para os dados de validação, comparado aquelas que foram obtidas na base de dados treino e teste com dados faltantes. Conclui-se que o modelo treinado com dados faltantes apresentou bom desempenho para os novos dados completos.

o para esse modelo e foi obtido os seguintes resultadosidaçe processamento do treinamento do modelo aumentava bastante.