Estudo Avaliativo de Modelos de Predição para baixa amostragem

Discente: Kaike Wesley Reis

Docente: Leizer Schnitman

Disciplina: ENGG11

I. OBJETIVO

O objetivo do trabalho prático II é avaliar diferentes abordagens de aprendizagem de máquina complementando o artigo de referência, para solucionar um problema de classificação de baixa amostragem no campo da saúde. Além disso, busca-se apresentar uma nova forma para diminuir a dimensionalidade do banco de dados.

II. INTRODUÇÃO

No período atual, aprendizagem de máquinas vem se tornando um tema de aplicação multidisciplinar. Sua utilização é praticamente universal para qualquer tipo de problema e suas vantagens são imensuráveis. Com isso torna-se fácil ver modelos aplicados tanto na biologia quanto na engenharia, passando por setores de saúde, marketing e etc.

Entretanto, alguns campos de estudos apresentam uma quantidade de amostras muito escassa para avaliar determinada problemática. Isso ocorre principalmente na área biológica em geral como na microbiologia, onde é comum possuir de 100 a 500 amostras para avaliar um estudo. A baixa amostragem pode se tornar um problema para aplicação de modelos preditivos, visto que eles acabam necessitando de um grande conjunto de informações para aprender os padrões e conquistar confiança em seus resultados.

O estudo proposto para o trabalho prático II irá avaliar como diferentes abordagens de inteligência artificial funcionam com um baixo número de amostragem e será apresentado uma nova forma para reduzir a dimensionalidade dos conjuntos de dados. Esse trabalho pode ser considerado um complemento ao artigo de referência que se encontra no tópico *Referências*.

O banco de dados estudado é referente ao câncer de mama. Ele consiste de 116 amostras onde 64 são pertencentes a pacientes que foram diagnosticados positivamente com câncer e 52 são pacientes de controle e portanto não apresentam a doença. Devido a baixa amostragem, alguns cuidados precisam ser tomados referente a confiança no modelo visto que o mesmo pode ficar sobre ajustado a aquele conjunto pequeno ou simplesmente não apresentar resultados coerentes de generalização.

As metodologias escolhidas foram:

K-Vizinhos Próximos (KNN) Trata-se de um algoritmo não supervisionado (não existe a necessidade de uma variável preditora) e não paramétrico (não necessita de uma definição especifica de distribuição para as variáveis, ou caso exista não apresenta definido sua parametrização). É considerado um dos modelos de inteligência artificial mais simples, visto que seu treinamento é apenas guardar o conjunto de treino. Ele avalia as amostras através das variáveis de entrada criando um espaço n-dimensional onde n é quantidade de features presente. Uma nova amostra é classificada de acordo aos K de vizinhos que definem seu agrupamento através de uma ponderação de distância no espaço. Seus principais hiper parâmetros são:

- K: Indica a quantidade de vizinhos que serão utilizados para definir o agrupamento de uma determinada amostra nova;
- Tipo de Distância: Como será calculado a distância entre os vizinhos e a nova amostra. O exemplo mais conhecido de distância é a euclidiana;
- Ponderação de Distância: Implica em como determinada distância vai afetar na decisão de classe da nova amostra. Um exemplo seria ponderar o resultado através da própria distância, ou seja, quanto mais próximo um vizinho for da amostra mais impacto ele terá na decisão em contra partida quanto mais longe for um vizinho menos impacto ele terá.

Rede Neural Artificial (ANN) É um modelo supervisionado (necessita de uma variável preditora) inspirado vagamente na constituição neurológica do cérebro animal constituido de neurônios e camadas (cada camada é um conjunto de neurônios). Devido a crescente expansão de estudos nessa área esse modelo apresenta diversos tipos diferentes de abordagem. Todavía, a abordagem avaliada aqui será a totalmente conectado. Ela caracteriza-se pelo total conexão dos neurônios de uma camada entre os neurônios da próxima camada. É uma ótima forma de aprendizagem de máquina podendo aproximar ou predizer diversos tipos de problemas, entretanto apresenta algumas imperfeições como o efeito caixa preta e a quantidade extensa de parâmetros a serem definidos. Seus hiper parâmetros principais são:

- Quantidade de camadas internas: Valor que identifica quantas camadas existirão no modelo. Esse valor como os outros não apresenta nenhuma base científica por trás e portanto é definida de acordo ao criador do modelo;
- Quantidade de neurônios em cada camada: Valores que identificam quantos neurônios irão existir em cada camada escondida e na camada de output;
- Tipo de treinamento: Como será dado o treinamento da rede. Atualmente existem diversas abordagens inspiradas no *back propagation*, que em resumo é uma propagação do erro gerado na resposta obtida para os pesos de conexão da rede.
- Função de saída dos neurônios das camadas internas e de saída: Função definida para a resposta dada pelo neurônios. Existe diversos tipos de funções, entretanto boa parte delas forçam com que a saída saiam em um intervalo fechado de zero a um.
- Função de Custo: Função definida para avaliar o erro do modelo que será propagado para os pesos da rede. A função mais comum é o erro quadrático.

Florestas Aleatórias (RF) é um modelo supervisionado caracterizado pela combinação de árvores de decisão, o que origina seu nome de florestas. Sua natureza aleatória advém de como ela é gerada: cada árvore escolhe de forma aleatória um subconjunto de features do conjunto maior para criar sua estrutura singular. A predição final do modelo é dado através de uma ponderação de todas as respostas de todas as árvores criando assim um modelo mais robusto e menos suscetível a questões de sobre ajuste (dado a aleatoriedade em que ele é criado). Esse modelo apresenta diversos hiper parâmetros, entretanto para esse estudo optou-se deixar a maioria no formato default. Os hiper parâmetros avaliados foram:

- Número de árvores: Valor que indica o número de árvores construídas pelo algoritmo antes de fazer uma votação ou uma média de predições. Em geral, uma quantidade elevada de árvores aumenta a performance e torna as predições mais estáveis, mas também torna a computação mais lenta;
- Fração InBag: Fração de amostras coletadas do banco de dados de treinamento com reposição;
- Número mínimo de folhas: Indica o número mínimo de folhas (nó que não apresenta conexão) que devem existir em uma dada árvore.

Bayes Inocente (NBC) é um classificador probabilístico que originou-se do teorema de bayes para gerar predições e recebe o nome de Inocente visto o pressuposto assumido pelo modelo de independência entre as variáveis de entrada, ou seja, ele assume que não existe nenhuma relação entre elas. Por ser um modelo probabilístico, existe a necessidade de definir ou assumir um tipo de distribuição para as variáveis de entradas para e avaliar os resultados, cálculos de verossimilhança e etc. Os hiper parâmetros avaliados foram:

- Nome da distribuição: Avalia o tipo de distribuição que melhor configura o conjunto de informações;
- Largura: Referente a quantidade de possíveis combinações entre um valor **x** de classes e **y** preditores no que se refere a uma distribuição do tipo Kernel;

- Tipo de Kernel: Dado a distribuição Kernel, é definido qual o tipo será referente a região de densidade.

Essas quatro abordagens serão analisadas para verificar qual apresenta melhor resultado na solução do problema proposto: criação de um modelo preditivo para a classificação da ausência ou presença de um tumor de mama dado a baixa amostragem.

Devido ao baixo conjunto amostral, é necessário aplicar uma técnica de redução de dimensionalidade. O estudo do artigo de referência avaliou diversas combinações entre as variáveis, escolhendo o melhor subconjunto como aquele que apresentasse o melhor modelo gerado. No trabalho prático II será proposto uma abordagem diferente através da análise de componentes principais.

A análise de componentes principais (PCA) é um procedimento matemático sensível a escala das variáveis (portanto é necessário transforma-las, a transformação mais comum é tornar a média zero e o desvio um) que utiliza uma transformação linear (ortogonal) para converter um conjunto de observações de variáveis possivelmente correlacionadas entre si em um conjunto de valores linearmente não correlacionadas e portanto independentes entre si chamados de componentes principais. Esta transformação é definida de forma que o primeiro componente principal tem a maior variância possível (ou seja, é responsável pelo máximo de variabilidade nos dados), e cada componente seguinte, por sua vez, tem a máxima variância sob a restrição de ser independente dos componentes anteriores. O toolbox do matlab gera diversos resultados que são referente a a análise feita, mas os principais são: os scores que trata-se basicamente dos componentes principais gerados e os loadings que representa o impacto das variáveis originais em cada componente principal.

III. METODOLOGIA DESENVOLVIDA

Será apresentado a metodologia para o trabalho prático II. Ela consiste em três grandes etapas: Pré-processamento, Modelos e Resultados. Cada uma será visualizada através de um fluxograma, apresentando suas partes e ao final um comentário geral a respeito do que foi criado. Caso seja necessário, será dado uma explicação mais aprofundada da parte de acordo a complexidade presente.

- **PARTE: Pré-processamento**

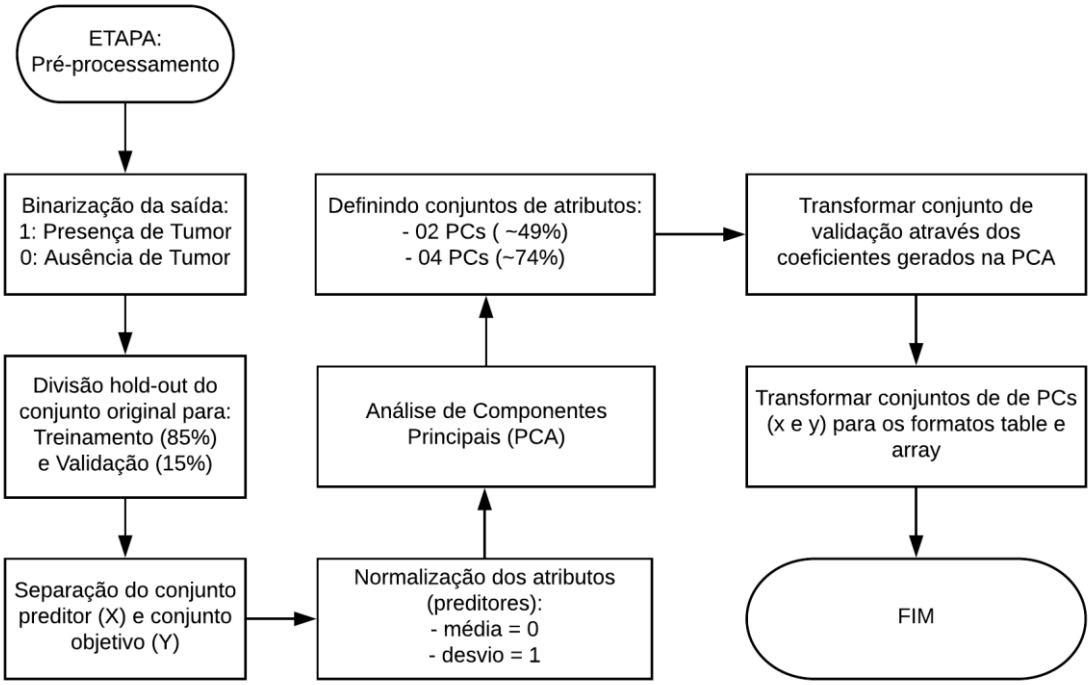


Figura 1: Fluxograma da parte Pré-processamento

O processo desenvolvido na parte Pré-processamento é semelhante ao do primeiro trabalho prático dessa disciplina. Trata-se de uma separação dos dados em treinamento e validação, análise de Componentes Principais e criação de um novo banco de dado a partir dessa análise, utilizando os componentes principais. É feito ao final uma transformação dos dados para obter o formato table e array no matlab evitando erros de compatibilidade com as funções disponíveis.

- **PARTE: Modelos**

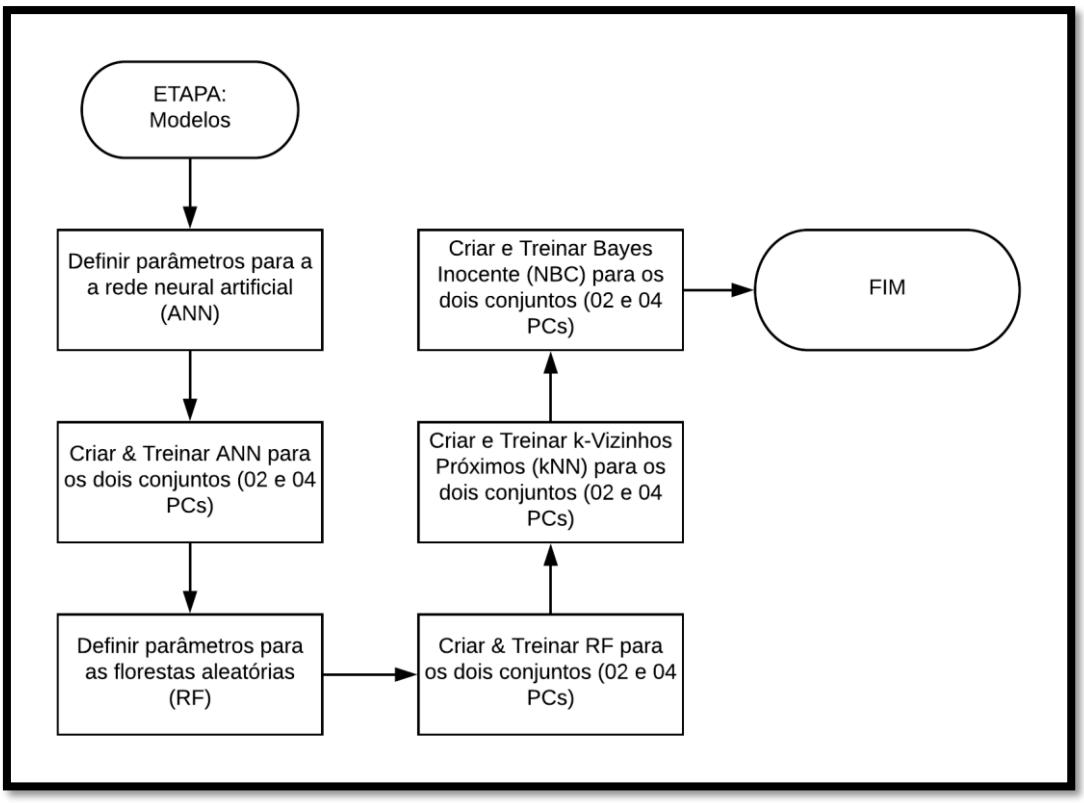


Figura 2: Fluxograma da parte Modelos

A parte Modelos é bem direta graças ao auxílio das ferramentas no matlab. Para os modelos k-NN e Bayes Inocente foi avaliado os melhores parâmetros no modelo para responder da melhor maneiro possível aquele banco de dados, ou seja, os parâmetros são otimizados para a problemática em questão. Para as Redes Neurais Artificiais e as Florestas Aleatórias é necessário definir os parâmetros de forma manual.

Abaixo são apresentados os parâmetros definidos de cada metodologia de predição:

- **Rede Neural Artificial:**
  - Camadas: (02 PCs) 2 camadas (04 PCs) 4 camadas
  - Neurônios por camada: (02 PCs) [10, 8] (04 PCs) [10, 8, 6, 4]
  - Treinamento: Levemberb-Marquardt
  - Função de output dos neurônios: Tangente Hiperbólica Sigmoidal
  - Função de output da última camada: Log-Sigmoidal
  - Função de Perda: Erro quadrático médio

- **K-Vizinhos Próximos:**
  - Tipo de distância: (02 PCs) Cityblock (04 PCs) Mahalanobis
  - Forma de Ponderação da distância: Quadrado inverso
  - Número de Vizinhos: (02 PCs) 3 Vizinhos (04 PCs) 14 Vizinhos

- **Bayes Inocente:**
  - Tipo de Distribuição: (02 PCs) Kernel (04 PCs) Normal
  - Largura: (02 PCs) 0.074 (04 PCs) NaN
  - Tipo de Kernel: (02 PCs) Triângulo (04 PCs) NaN

- **Florestas Aleatórias:**
  - Número de Árvores: 1000
  - Fração *InBag*: 1
  - Número minímo de folhas: 1

- **PARTE: Resultados**

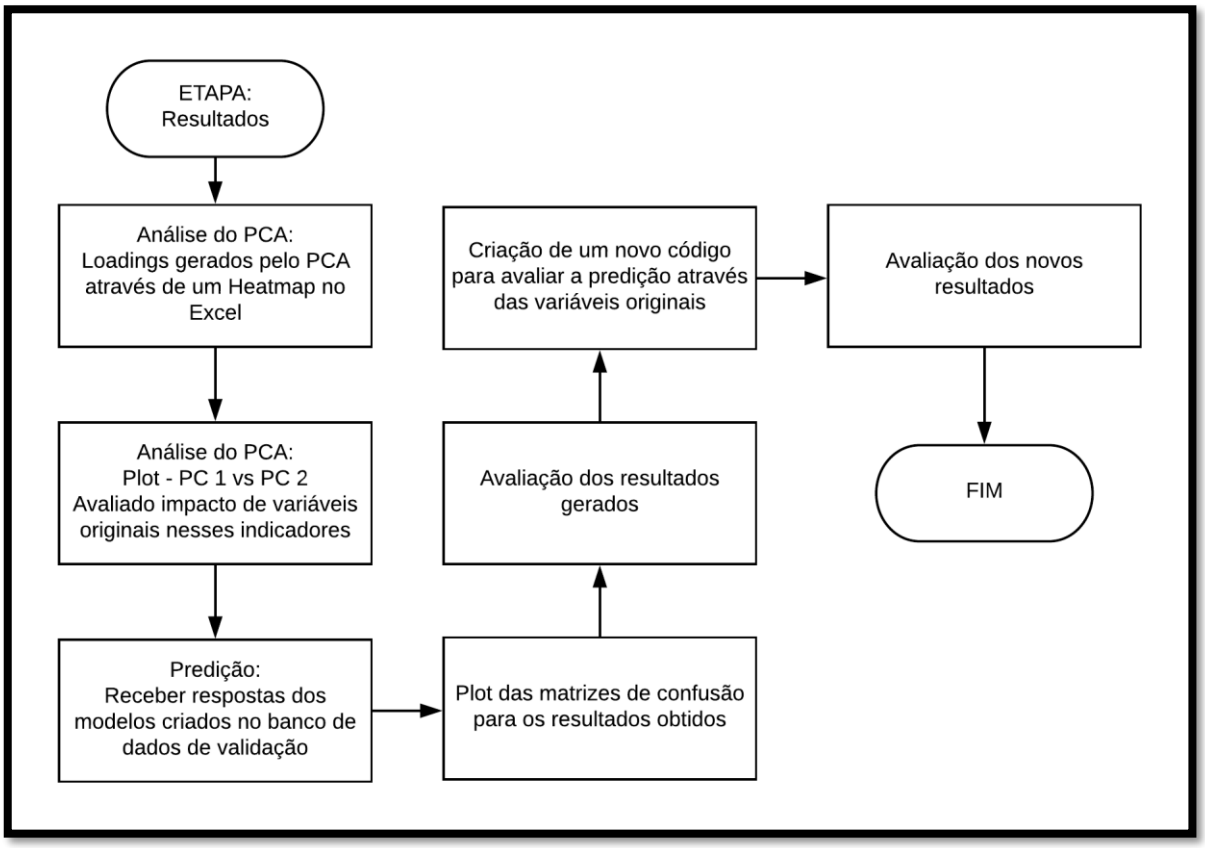


Figura 3: Fluxograma da parte Resultados

A parte Resultados avalia a análise de componentes principais e os resultados de predições dos modelos gerados. Logo em seguida, é desenvolvido um código semelhante ao que foi desenvolvido até esse ponto com o foco em avaliar as mesmas formas de aprendizado de máquina utilizando como preditoras as variáveis escolhidas como boas preditoras no artigo de referência.

IV. RESULTADOS

Três resultados foram gerados inicialmente: mapa de calor para a análise de componentes principais, gráfico dos componentes principais e as métricas de predições acurácia, sensibilidade e especificidade. Ao final, notou-se que a PCA pode não ter sido uma boa abordagem para redução de dimensionadalidade. Então foi criado um segundo código para avaliar as metodotologias de aprendizagem de máquina utilizando agora como entrada as quatro variáveis escolhidas pelo artigo de referência como boas preditoras para a problemática. O intuito é avaliar se as variáveis originais podem ser features mais consistentes para apresentar modelos com métricas melhores.

- **Mapa de Calor para a Análise de Componentes Principais**

O mapa de calor para a análise de componentes principais trata-se de uma tabela entre as variáveis originais e os componentes principais escolhidos para compor o conjunto de entradas durante o trabalho prático.

Os valores apresentados são os *loadings* gerados pela PCA e sua interpretação é dada seguinte forma: quanto maior o módulo, mais impacto aquela variável possui para alterar determinado componente principal. O sinal implica a relação entre eles sendo proporcional ou inversamente proporcional caso seja positivo ou negativo respectivamente.

O mapa de calor é apresentado abaixo:

ORIGINAIS	PC1	PC2	PC3	PC4
Age	0,0672	0,012	-0,3212	0,8366
BMI	0,2953	0,463	-0,3085	-0,2477
Glucose	0,4299	-0,1717	0,0419	0,2394
Insulin	0,4494	-0,4062	-0,0361	-0,1573
HOMA	0,4985	-0,4027	0,015	-0,0417
Leptin	0,3155	0,305	-0,552	-0,212
Adiponectin	-0,1963	-0,3778	-0,1342	-0,3304
Resistin	0,2347	0,3884	0,369	-0,0026
MCP1	0,2826	0,2053	0,5826	0,0288

Tabela 1: Mapa de calor gerado para os quatro primeiros componentes principais.

A coloração dos *loadings* está relacionado com o módulo e com o tipo de relação. Quanto maior o módulo mais forte será a tonalidade e as cores verde e vermelha indicam respectivamente relações proporcionais e inversamente proporcionais. Vale destacar que as variáveis que constitui o conjunto de features do melhor modelo no artigo de referência estão grifadas em vermelho.

Através de sua análise, verifica-se que a PCA não conseguiu distinguir uma importância relevante entre as variáveis, devido ao fato de que todas as variáveis possuem ao menos duas relações consideráveis (com alguma coloração) com os componentes estudados. Isso difere do artigo de referência que apresentou quatro possíveis preditoras que se destacam na performance dos modelos gerados por elas.

Avaliando, por exemplo, os dois primeiro componentes principais, verifica-se que de nove variáveis apenas uma não possui um impacto relevante em ambas (no caso, Age) e os loadings não se diferenciam tanto em módulo.

Um ponto que vale ressaltar, é que a variável BMI apresentada como uma das escolhidas no estudo de referência também é apresentada na PCA como uma variável importante visto que ela impacta em todos os componentes principais destacados (juntos, eles explicam aproximadamente 74% da variância dos dados), isso pode indicar que de fato essa variável possui aspectos relevantes para o objetivo estudado no artigo.

- **Gráfico dos Componentes Principais**

O gráfico se trata de um plano entre componentes principais, os escolhidos foram os que apresentam maior explicação de variância dos dados, onde o primeiro e o segundo componentes explicam respectivamente 34% e 15% da variância total do conjunto original. É plotado as varíaveis originais, que foram transformadas para o domínio de componentes através dos *loadings*. O gráfico é apresentado abaixo:

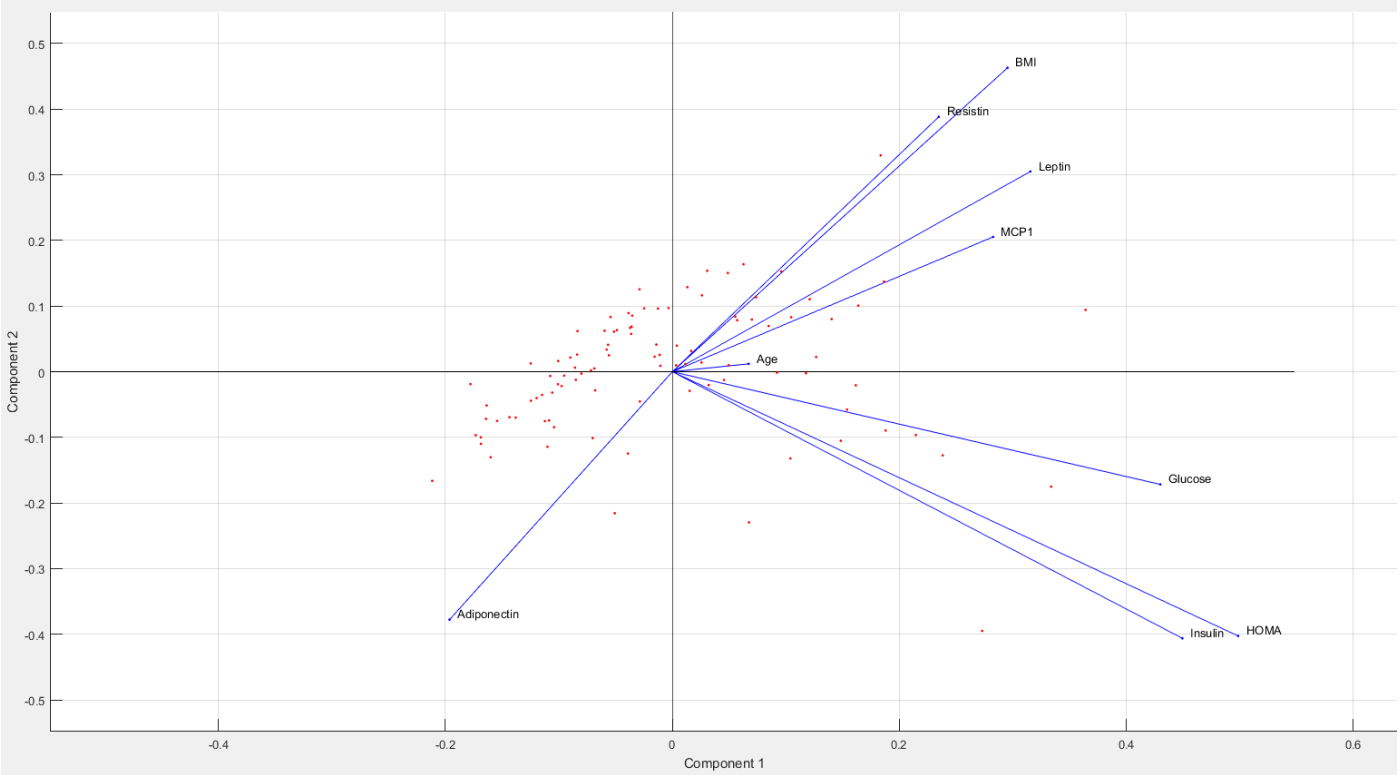


Figura 4: Gráfico de Componente Principal 1 vs Componente Principal 2

Verifica-se que esse gráfico se trata de uma extensão do mapa de calor apresentado anteriormente. Por exemplo, as variáveis Insulin e HOMA apresentam para o primeiro componente principal loadings de maior módulo para tendência positiva e para o segundo componente loadings de maior módulo com tendência negativa, portanto suas retas se destacam em comprimento e localizam-se no terceiro quadrante. A variável Age que apresenta loadings inexpressivos e positivos no mapa de calor, encontra-se no primeiro quadrante com uma reta pequena.

Partindo dessa premissa, é possível avaliar de forma gráfica o impacto das variáveis na definição dos componentes principais de acordo a sua localização e módulo. Esse resultado é confirmado através da avaliação do mapa de calor gerado anteriormente e traz consigo implicações referente a importância das variáveis para o problema estudado.

- **Métricas dos modelos**

No total, foram avaliados oito modelos que utilizaram quatro abordagens de aprendizagem de máquina em dois diferentes bancos de dados (dois e quatro componentes principais). As métricas podem ser entendidas da seguinte forma: especificidade indica a porcentagem que o modelo previu corretamente a ausência do tumor, sensibilidade indica a porcentagem que o modelo previu corretamente a presença do tumor e a acurácia avalia o quanto o modelo acertou de modo geral. O nome dos modelos são dados de acordo a metodologia aplicada através de siglas, K-Vizinhos Próximos, Rede Neural Artificial, Bayes Inocente e Florestas Aleatórias são representadas respectivamente por K-NN, ANN, NBC, RF; O nome é completado de acordo ao banco de dados utilizado variando entre dois e quatro componentes principais.

A tabela abaixo resume as matrizes de confusões geradas por cada modelo (em anexo 1 até 9 é possível encontrar uma explicação sobre o siginificado de cada quadrante da matriz de confusão gerada pelo matlab seguido pelas matrizes de confusões para todos os modelos) ao avaliar o banco de dados de validação:

MODELOS \ MÉTRICAS	ESPECIFICIDADE %	SENSIBILIDADE %	ACURÁCIA %
K-NN 02 PCs	57.1	80.0	70.6
K-NN 04 PCs	71.4	50.0	58.8
ANN 02 PCs	28.6	90.0	64.7
ANN 04 PCs	71.4	50.0	58.8
NBC 02 PCs	57.1	60.0	58.8
NBC 04 PCs	85.7	40.0	58.8
RF 02 PCs	71.4	70.0	70.6
RF 04 PCs	57.1	60.0	58.8

Tabela 2: Métricas avaliadas em cada modelo para os componentes principais

Nos resultados encontrados, é possível perceber que nenhum modelo performou de maneira excelente. Os modelos criados de um modo geral apresentam métricas semelhantes entre sí e medíocres. A maioria supera pouco um modelo randômico (métricas equivalente a 50%) e apenas um conseguiu alcançar resultados maiores que 70%. Esse resultado pode indicar que a PCA pode não ter sido a melhor ferramenta para diminuir a dimensionalidade desse conjunto de informações.

É possível identificar casos de sobre ajuste como no modelo *ANN 02 PCs* com uma sensibilidade de 90% entretanto sua especificadade é baixíssima, indicando uma tendência do modelo a sempre chutar em uma classe apenas. Isso pode ser verificado também com um grau menor no modelo *NBC 04 PCs*, que apresenta uma alta especificidade que é compensada por uma sensibilidade abaixo de 50%.

Os modelos que se destacaram foram *KNN 02 PCs* e *RF 02 PCs* visto que suas métricas podem ser consideradas razoáveis. Entretanto o segundo modelo apresenta mais consistência nas três métricas e portanto foi escolhido como o melhor modelo entre os oito gerados. Todavía, não é suficiente para ser comparado com o melhor modelo proposto pelo artigo de referência.

- **Modelos aplicados nas variáveis escolhidas pelo artigo de referência**

Dado o baixo desempenho dos modelos de uma forma geral utilizando a PCA, foi avaliado as mesmas metodologias com os mesmos parâmetros para Redes neurais artificiais e florestas aleatórias e para k-Vizinhos Próximos e Bayes Inocente utilizou-se novamente a função de otimização do toolbox para encontrar os melhores parâmetros. Os modelos foram então treinados com as quatro variáveis apresentadas no artigo de referência como conjunto de boas preditoras, responsáveis por definir o melhor modelo.

Os parâmetros encontrados foram:

- **K-Vizinhos Próximos:**
  - Tipo de distância: seuclidian
  - Forma de Ponderação da distância: Quadrado inverso
  - Número de Vizinhos: 31 Vizinhos

- **Bayes Inocente:**
  - Tipo de Distribuição: Normal

- Largura: NaN
- Tipo de Kernel: NaN

Os resultados preditivos foram sumarizados na tabela abaixo (em anexo 10 encontra-se a matriz de confusão do melhor modelo, a rede neural artificial):

MODELOS \ MÉTRICAS	ESPECIFICIDADE %	SENSIBILIDADE %	ACURÁCIA %
K-NN	57.1	80.0	70.6
ANN	85.7	90.0	88.2
NBC	85.7	60.0	70.6
RF	85.7	80.0	82.4

Tabela 3: Métricas avaliadas em cada modelo para as variáveis originais escolhidas

Nota-se claramente uma melhora. Verifica-se que o melhor modelo agora são as redes neurais com métricas excelentes seguido pela florestas aleatórias. K-Vizinhos Próximos e Bayes Inocente apesar de melhorarem suas métricas ainda não configuraram um desempenho satisfatório comparado aos dois primeiros. Diversas explicações podem ser dadas para esse baixo desempenho, uma delas seria que tanto K-NN quanto NBC não são boas abordagens para tal problemática. Os resultados expostos, podem ser comparados com o modelo apresentado como o melhor no artigo de referência devido ao aumento das métricas, entretanto existe algumas ressalvas que serão explicadas durante a conclusão.

V. COMPARAÇÕES COM ESTUDO REFERÊNCIA

O artigo utilizado como referência para esse projeto apresenta uma metodologia semelhante. Os autores avaliaram três métodos de aprendizagem de máquinas: Regressão Logística, Florestas Aleatórias e Máquina de Vetores de Suporte através de um processo de validação cruzada, utilizando como preditoras diversas combinações das nove variáveis existentes no banco de dados.

O modelo escolhido seria aquele que apresentasse as maiores métricas com um intervalo de confiança: Área abaixo da curva (AUC) associada a análise da curva de Características Operacionais de Repector (ROC), Sensibilidade e Especificidade.

O melhor modelo foi uma Máquina de Vetores de Suporte que utilizou como preditoras as variáveis Age, BMI, Glucose e Resistin. As métricas alcançadas se encontram abaixo:

- Sensibilidade - [82%, 88%]
- Especificidade – [84%, 90%]
- AUC – [0.87, 0.91]

Vale ressaltar que para o tipo de problemática, a sensibilidade apresenta um peso maior visto que ela identifica o erro de um diagnóstico de cura, quando na verdade o paciente apresenta o tumor.

O trabalho prático aqui desenvolvido apresenta dois modelos com algumas métricas melhores ou equivalentes ao apresentado no artigo, vale lembrar que não é utilizado a métrica AUC e portanto será apenas avaliado as duas restantes. A rede neural artificial gerada a partir das mesmas variáveis que criaram o modelo escolhido apresenta uma sensibilidade maior que o intervalo de confiança e uma especificidade contida no intervalo de confiança do modelo escolhido. Para a floresta aleatória criada a partir do conjunto de boas preditoras, segundo melhor modelo desse trabalho, apresenta o uma especifidade dentro do intervalo de confiança proposto e uma

sensibilidade abaixo do intervalo de confiança. Nenhum modelo criado a partir de componentes principais alcançou resultados desejados para serem avaliados.

VI. CONCLUSÃO

O objetivo desse trabalho prático foi alcançado. Foi apresentado uma avaliação de novos novos modelos para solucionar a problemática, uma abordagem diferente para tratar a redução de dimensionalidade (Análise de Componentes Principais).

Vale ressaltar que através de todos os resultados demonstrado neste trabalho prático, a PCA pode não ser considerada uma boa solução para reduzir a dimensionalidade das variáveis de entradas, portanto conclusões feitas a partir dela devem ser descartadas. Isso pode ter ocorrido principalmente devido ao conjunto de dados não ser normalmente distribuído conjuntamente; A não validação dessa prerrogativa não garante que os componentes principais sejam totalmente independentes (isso ocasiona por exemplo, um erro de definição do classificador bayes inocente que tem em sua definição a afirmação de indenpedência entre as variáveis de entrada e portanto implica na baixa de suas métricas). Isso se torna mais claro ao avaliar que os modelos criados a partir de componentes principais não obtiveram boas métricas preditivas, sendo abaixo do razoável. Outro ponto de discordância seria os loadings dos componentes principais, que identificaram importância similares entre uma grande parcela das variáveis, o que diverge do estudo de referência que identifica quatro das noves variáveis como boas preditoras.

Além disso, é apresentado que as redes neurais artificiais podem ser uma opção considerável em estudos futuros visto as excelentes métricas alcançadas utilizando as mesmas variáveis preditoras do artigo de referência, porém essa metodologia deve ser usada com ressalvas.

Analisando com cautela informações sobre a rede neural criada podemos verificar:

- Avaliação de Performance durante o treinamento
- Histograma de Confiança

Abaixo, cada tópico acima será explicado com mais detalhes:

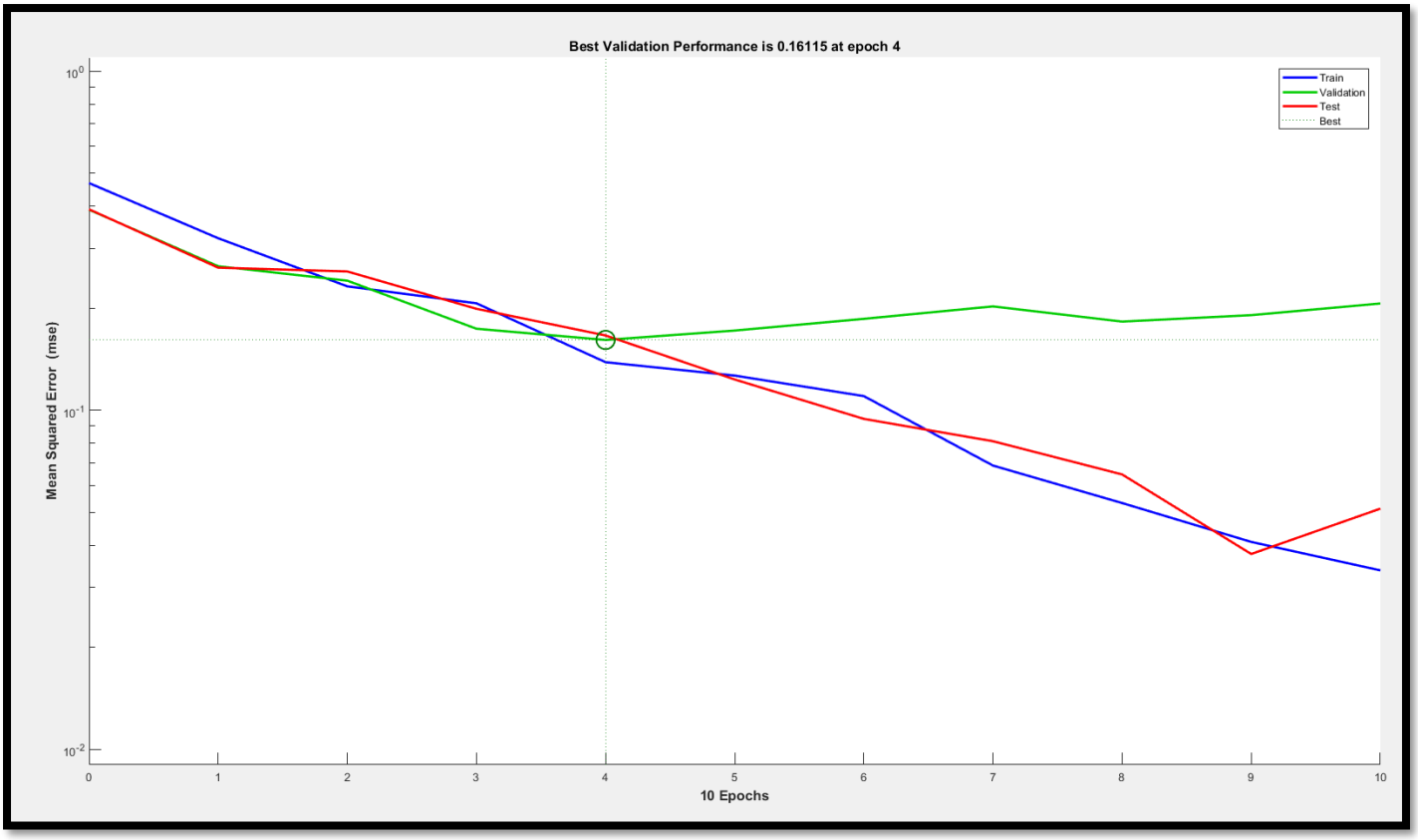


Figura 5: Avaliação de performance do treinamento do modelo

Inicialmente é possível verificar uma baixa quantidade de Epochs para treinar o modelo, isso é decorrente do formato de treinamento escolhido ser excessivamente mais rápido que qualquer outro presente no toolbox e a baixa quantidade amostras. As curvas mostram que tanto a parte de treinamento quanto testes decaem ao longo do tempo em relação ao erro quadrático (um resultado esperado, afinal um ótimo modelo tende a errar menos), entretanto a fase de validação começa a apresentar um erro crescente ao final, algo não esperado e que pode prejudicar a confiança nas métricas obtidas, pois isso pode indicar sobre ajuste do modelo.

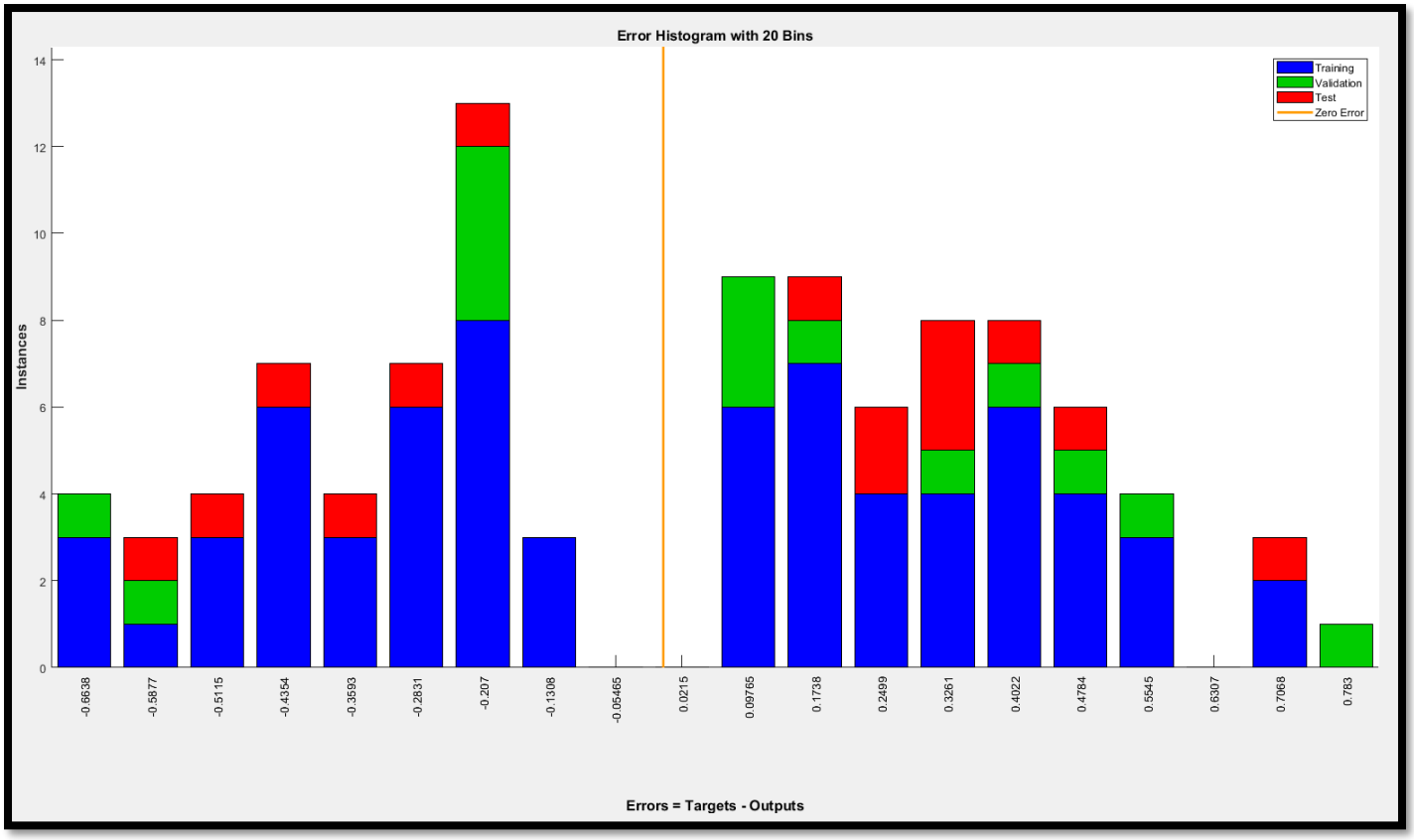


Figura 6: Histograma de Confiança da Rede Neural durante o treinamento

Nota-se que a rede neural apresenta respostas com um nível de confiança intermediário. Esse histograma pode ser avaliado da seguinte forma: do lado direito temos diferenças de respostas associadas a ausência do tumor (valor zero) e a esquerda a presença de tumor (valor um). É mostrado que a confiança do modelo gira em torno de 20% da 10% para predições em relação a ausência e presença do tumor respectivamente dado aos picos estarem em torno desses resultados. Entretanto, o histograma tende a apresentar valores em suas extremidades o que implica falta de confiança do modelo em suas respostas. Acredita-se que a baixa amostragem pode ser responsável por esse histograma, logo para tornar o modelo mais robusto pode ser necessário uma quantidade maior de amostras.

Vale ressaltar que a fase de validação apresenta valores baixos de confiança, a medida que ela ela está presente ao longo do histograma em direção as extremidades.

Portanto, através desses resultados é possível verificar que a redes neural, para apresentarem resultados mais robustos e de confiança necessita de um conjunto amostral maior. Logo, caso a problemática ainda consista em um conjunto baixo de amostragens é indicado o modelo apresentado pelo artigo de referência.

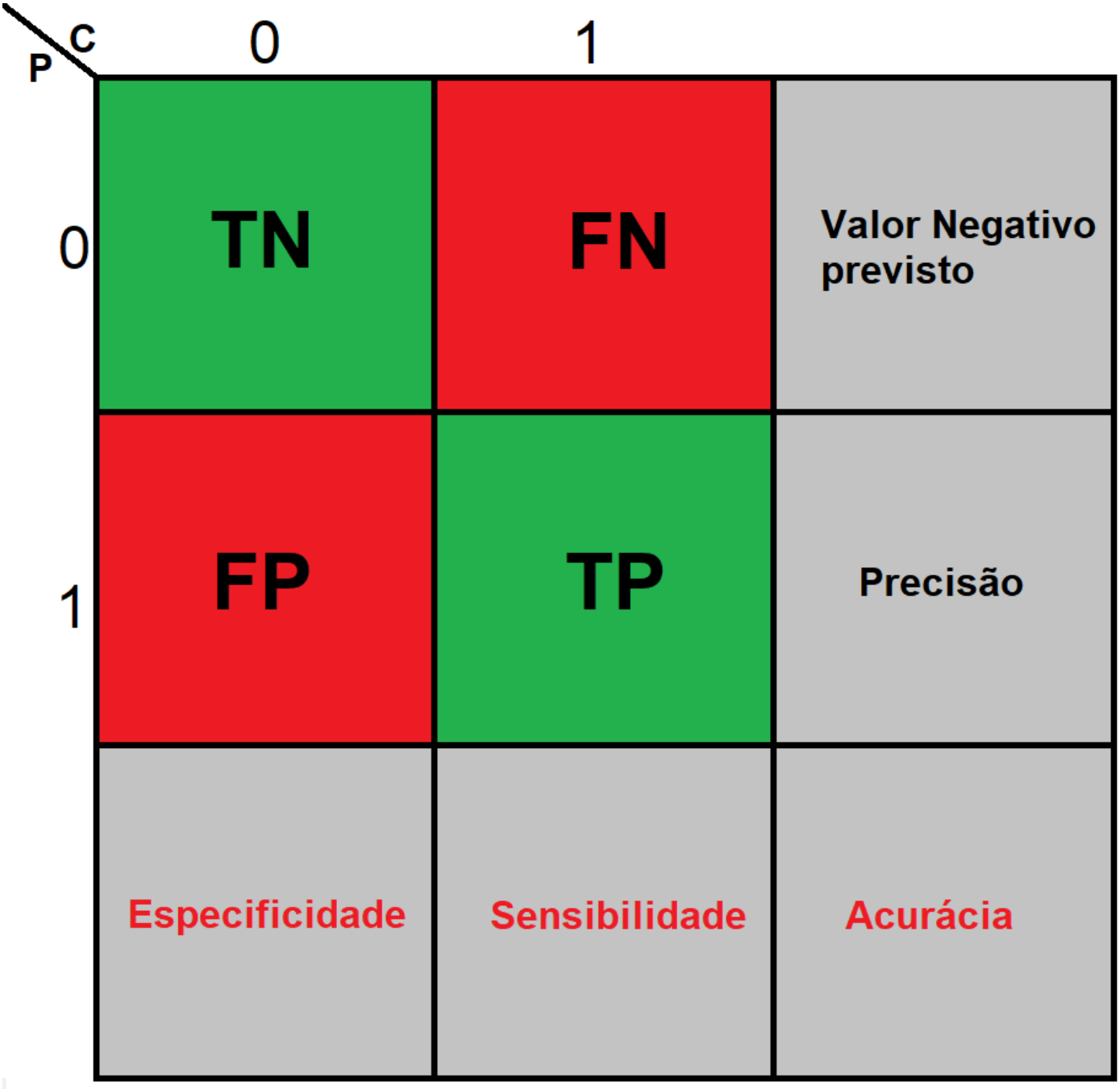
É possível propor após a finalização desse trabalho, um foco maior nas redes neurais artificiais, máquinas de vetores de suporte e florestas aleatórias (apesar de não se destacarem, tanto aqui quanto no estudo referência seus resultados são considerados bons em ambos). Técnicas de aumento de amostras como SMOTE (Synthetic Minority Over-sampling Technique) podem ser utilizadas para contornar a baixa quantidade de informações, isso viabiliza inclusive avaliar as outras metodologias já abordadas nesse estudo e no artigo, verificando se com um conjunto maior de amostras elas conseguem apresentar performances melhores.

VII. REFERÊNCIAS

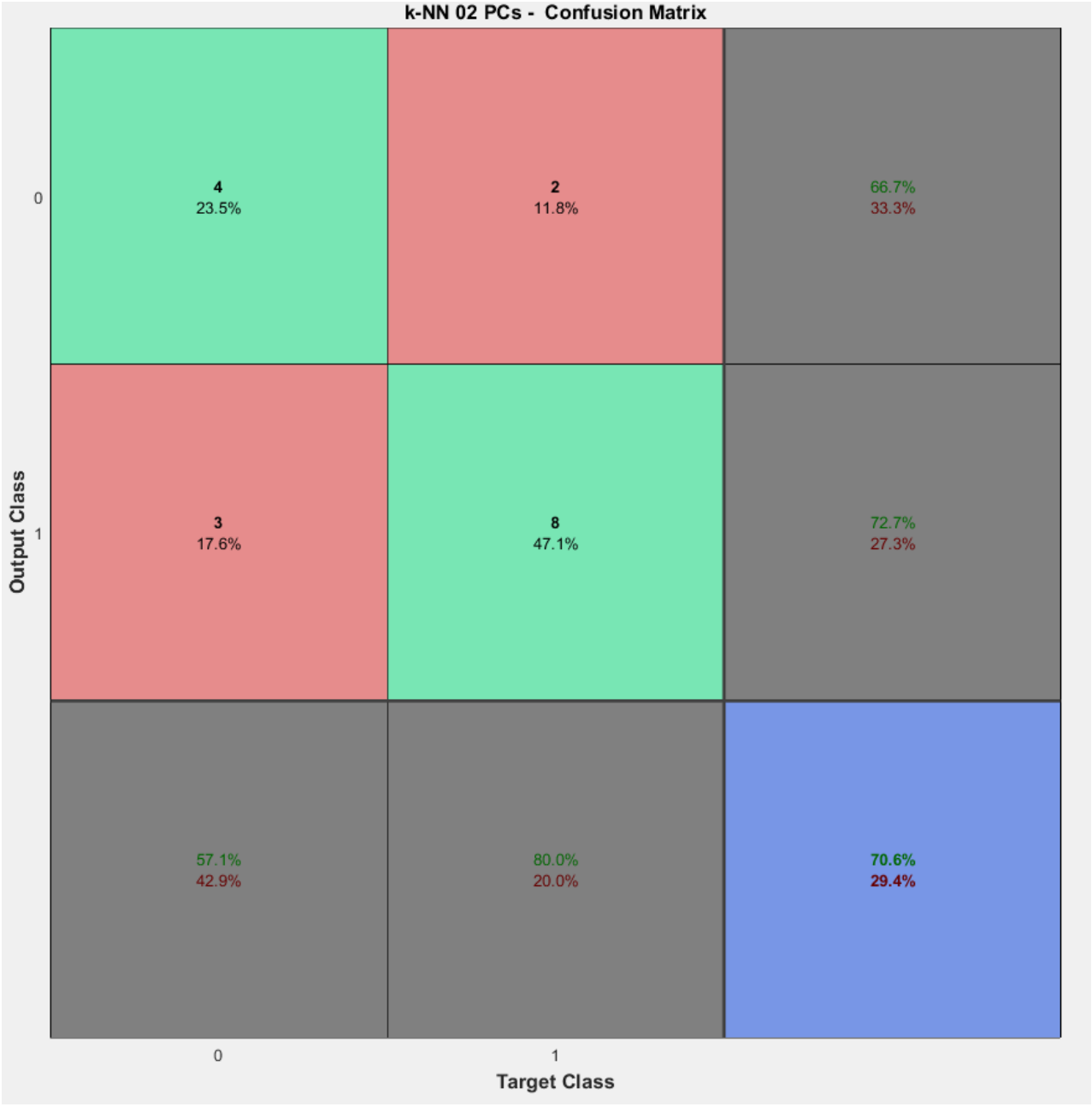
- Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F.; BMC Cancer; Using Resistin, glucose, age and BMI to predict the presence of breast câncer; 2018.

VIII. ANEXOS

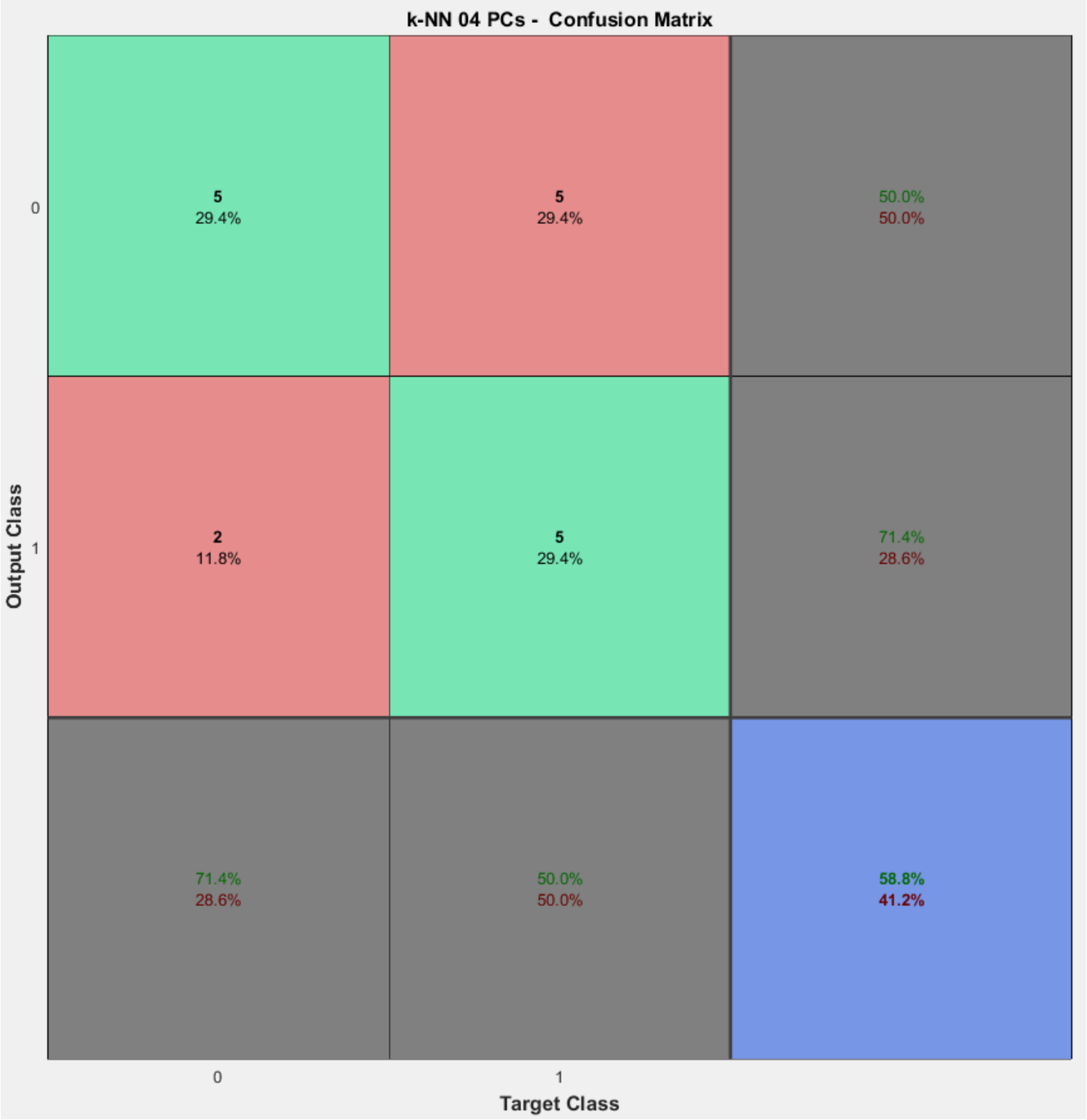
Abaixo encontra-se uma breve explicação de cada quadrante da matriz de confusão seguido pelas matrizes de confusões para os modelos criados a partir da PCA e finalizando com a matriz de confusão do modelo gerado pelo conjunto de boas preditoras:



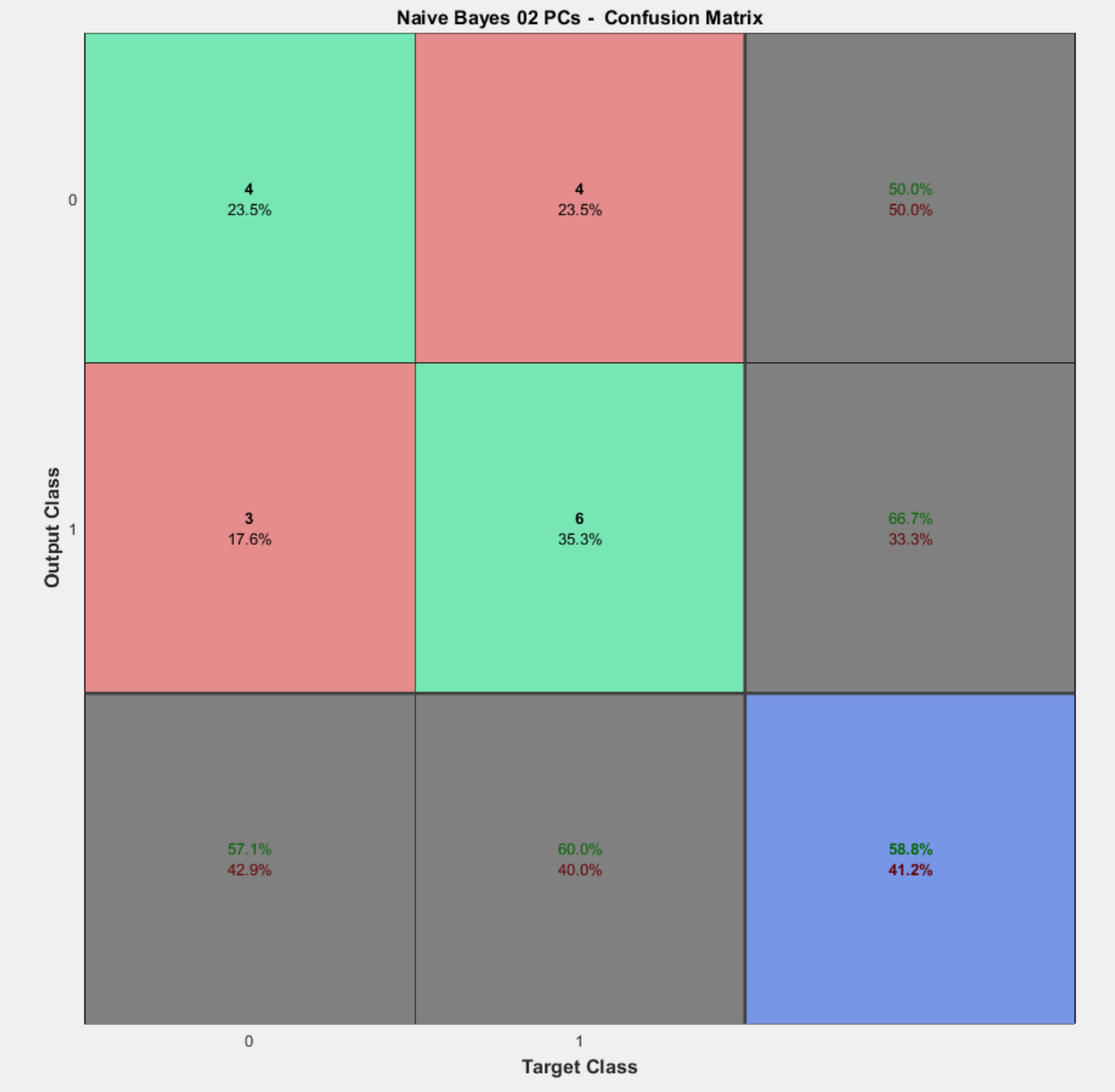
Anexo 1: Explicação de cada quadrante da matriz de Confusão gerada pelo Matlab. No canto superior esquerdo P implica em classe predita pelo modelo e C implica em classe correta dita pelo banco de dados



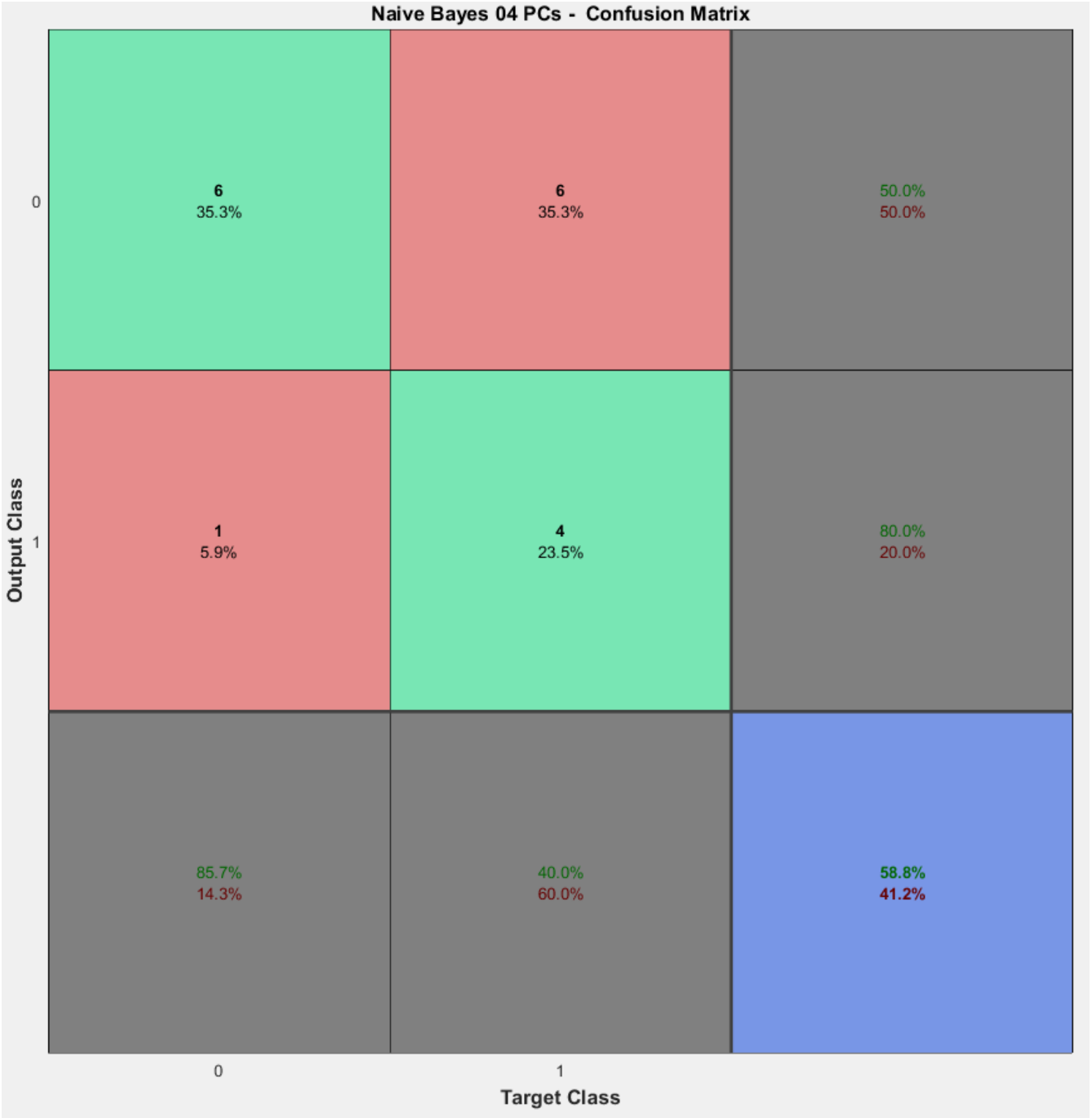
Anexo 2: Matriz de Confusão para o modelo K-NN com dois componentes principais



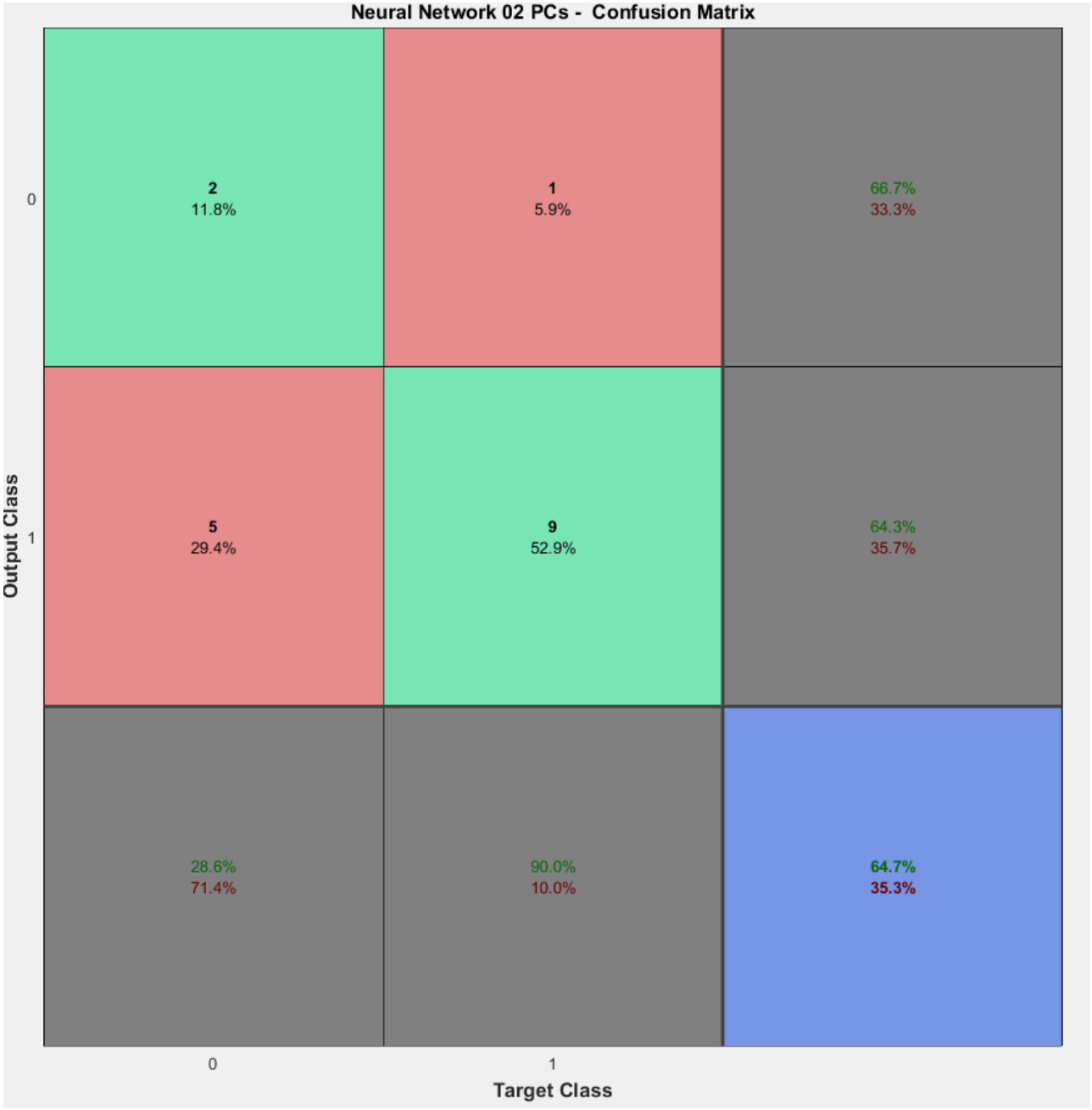
Anexo 3: Matriz de Confusão para o modelo K-NN com quatro componentes principais



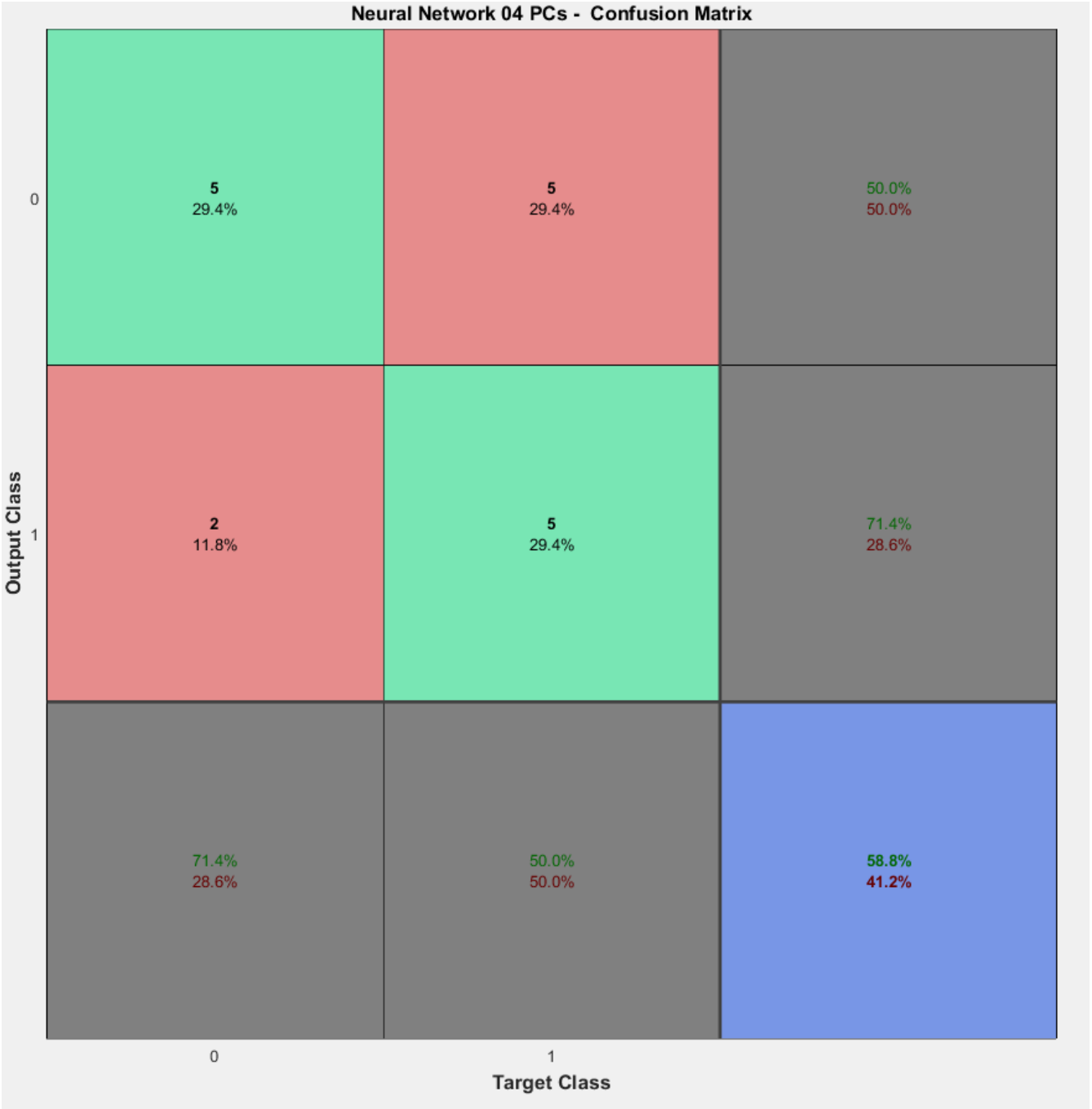
Anexo 4: Matriz de Confusão para o modelo Bayes Inocente com dois componentes principais



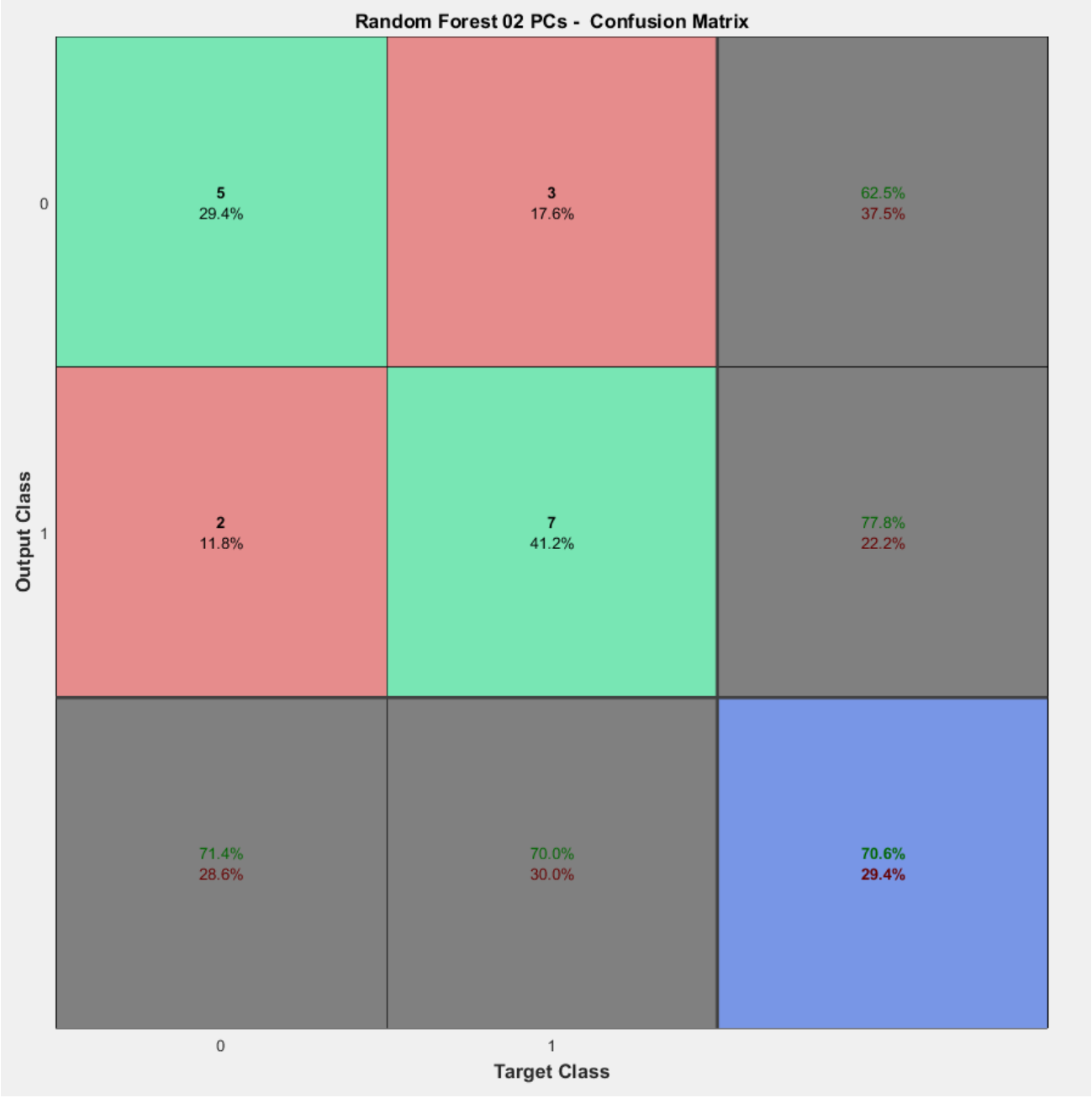
Anexo 5: Matriz de Confusão para o modelo Bayes Inocente com quatro componentes principais



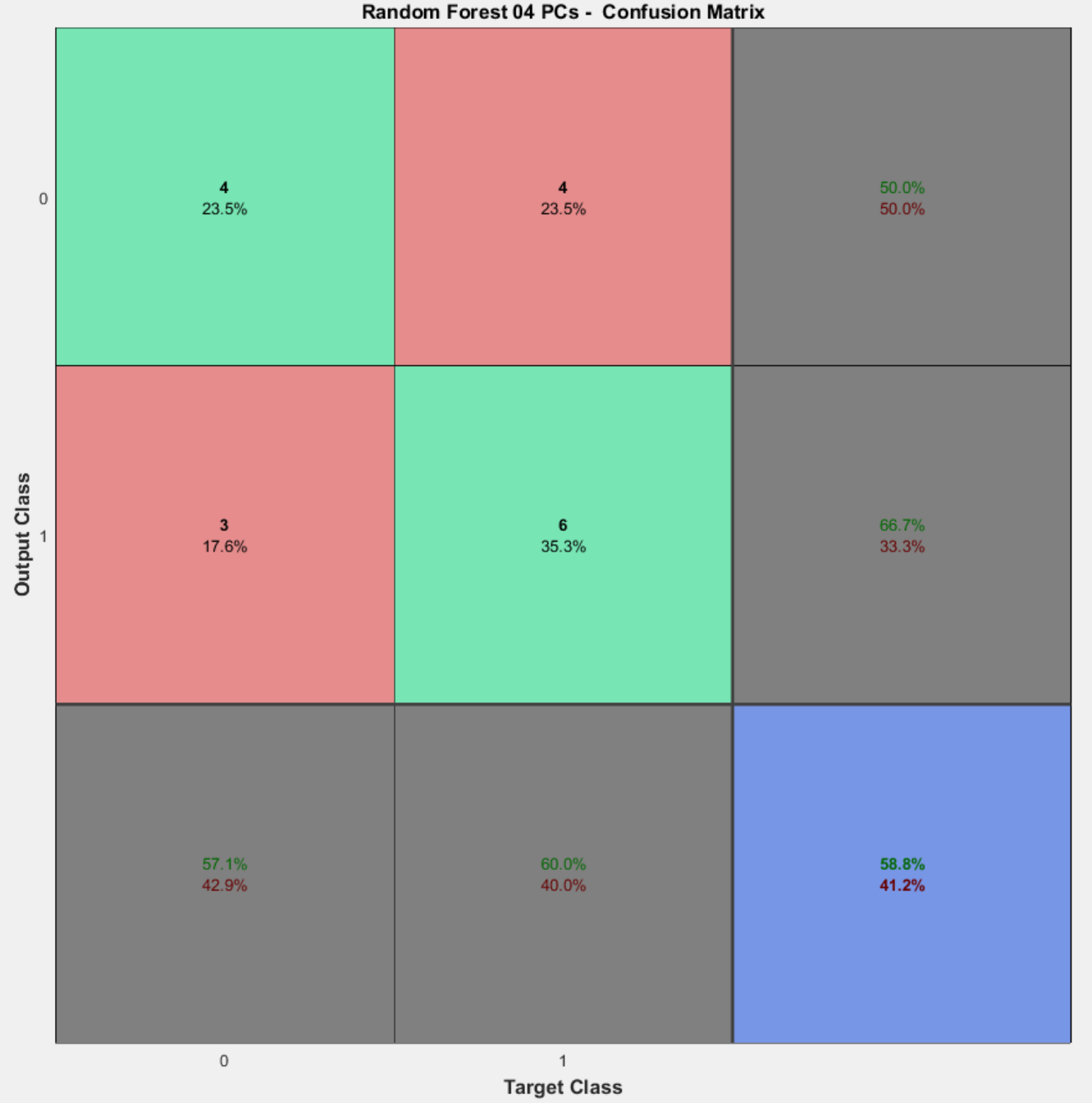
Anexo 6: Matriz de Confusão para o modelo Rede Neural com dois componentes principais



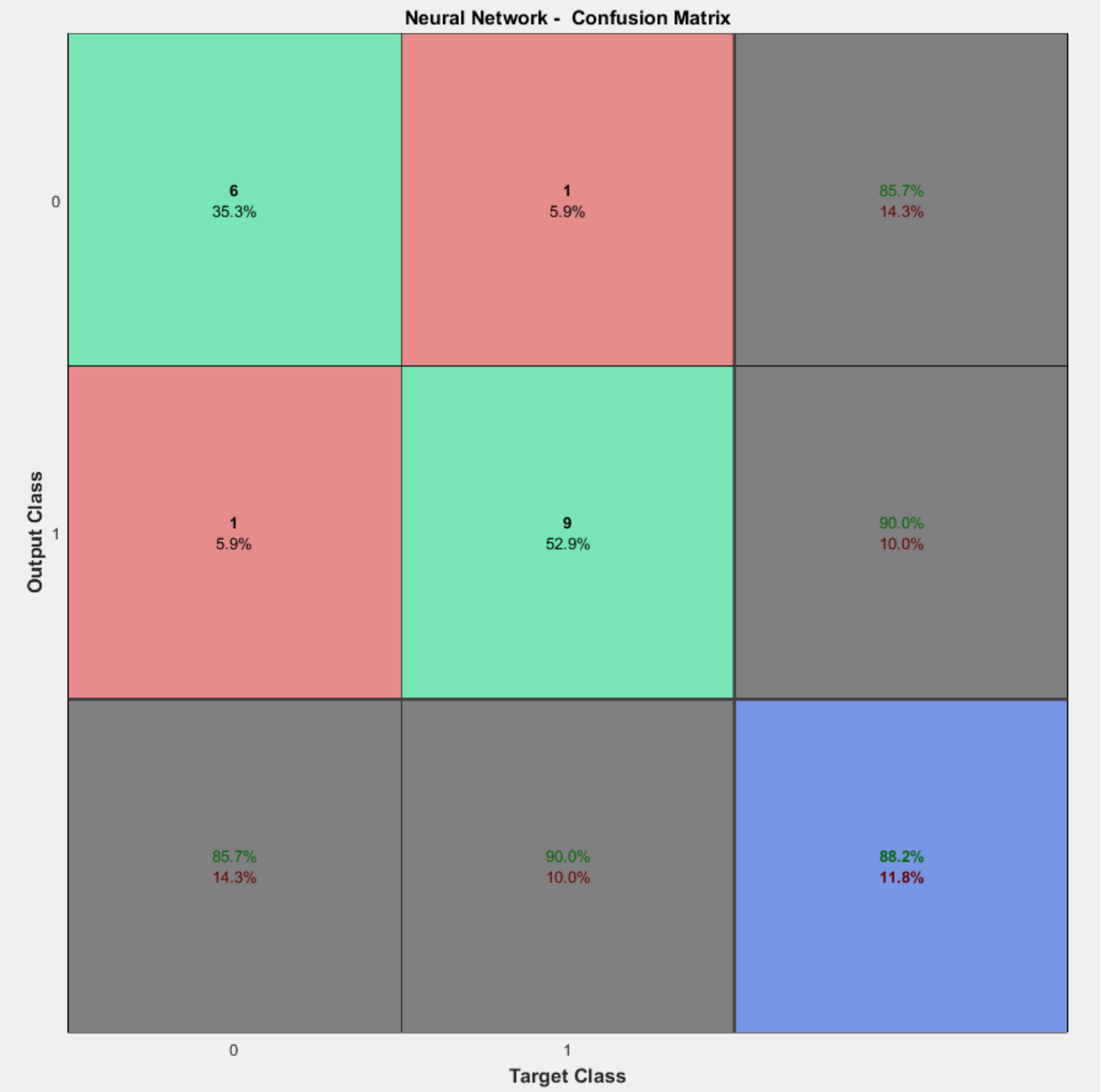
Anexo 7: Matriz de Confusão para o modelo Rede Neural com quatro componentes principais



Anexo 8: Matriz de Confusão para o modelo Florestas Aleatórias com dois componentes principais



Anexo 9: Matriz de Confusão para o modelo Florestas Aleatórias com quatro componentes principais



Anexo 10: Matriz de Confusão para o modelo Redes Neurais Artificias criado a partir do conjunto de variáveis originais dito pelo estudo de referência como melhores preditoras