

Universidade de São Paulo

Instituto de Ciências Matemáticas e de Computação

Projeto I

SCC 0270

Redes Neurais

Profa. Dra. Roseli Aparecida
Francelin Romero

Alunos

Moisés Botarro Ferraz Silva, 8504135

Thales de Lima Kobosighawa, 9897884

Victor Rozzatti Tornisiello, 9806867

Sumário

Introdução	3
Classificação	4
Descrição da Base de Dados	4
Pré-Processamento dos Dados	4
Normalização	4
Balanceamento	5
Treinamentos e Testes Realizados	5
Arquitetura da Rede	5
Número de Épocas	7
Taxa de Aprendizado e Momentum	8
Proporções nos Conjuntos de Treinamento e Teste	9
Undersampling e Oversampling da Base Original	9
Regressão	12
Descrição da Base de Dados	12
Pré-Processamento dos Dados	12
Normalização	12
Treinamentos e Testes Realizados	12
Arquitetura da Rede	12
Número de Épocas	13
Taxa de Aprendizado e Momentum	14
Proporção nos Conjuntos de Treinamento e Teste	15
Avaliação do Modelo Final	16

1. Introdução

No projeto realizado, implementamos um MultiLayer Perceptron e o utilizamos em um problema de Classificação e Regressão. A fim de melhor organizar o código, separamos esses dois problemas em dois arquivos Jupyter Notebook distintos: *Projeto 1 - Classificação* e *Projeto 1 - Regressão*.

Em cada um desses arquivos, descrevemos todas as etapas que foram realizadas para a construção do modelo que possibilita a obtenção dos melhores resultados para cada problema.

No relatório a seguir, apresentamos resumidamente as etapas seguidas na avaliação do impacto dos meta parâmetros dos modelos em seu desempenho, bem como os resultados obtidos.

2. Classificação

2.1. Descrição da Base de Dados

A base de dados fornecida para classificação contém os dados resultantes de uma análise físicoquímica realizada em vinhos vermelhos. Ela possui 11 atributos e 3 classes diferentes, sendo elas 'Bad', 'Good' e 'Mid', que dizem respeito à qualidade do vinho analisado.

Os dados são organizados em forma de tabela, onde cada linha corresponde a uma amostra e cada coluna corresponde a um atributo, além de uma primeira coluna extra para indicar o índice da amostra e uma última coluna extra para indicar sua classe. Os exemplos estão ordenados por classe, e não são balanceados, uma vez que existem muito mais vinhos considerados normais do que vinhos especialmente bons ou ruins. Além disso, os atributos não possuem a mesma ordem de grandeza.

Na figura 1 podem ser observadas as cinco primeiras amostras presentes na base de dados, sem qualquer tipo de pré-processamento realizado. No total, são 1599 exemplos, sendo 1319 pertencentes à classe 'Mid', 217 pertencentes à classe 'Good' e apenas 63 pertencentes à classe 'Bad'.

	Unnamed: 0	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	category
0	0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	Mid
1	1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	Mid
2	2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	Mid
3	3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	Mid
4	4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	Mid

Figura 1 - Algumas amostras da base de dados *Red Wine Quality*.

2.2. Pré-Processamento dos Dados

2.2.1. Normalização

Como primeiro pré-processamento realizado sobre os dados do conjunto, foi realizada a normalização dos valores de cada atributo, de forma que cada um deles se encontre no intervalo $(0,1]$. Isso foi feito para evitar a saturação na saída dos neurônios devido a valores de atributos muito altos e para melhorar a convergência do algoritmo de classificação.

2.2.2. Balanceamento

Além disso, conforme mencionado anteriormente, a base de dados é desbalanceada, possuindo apenas 63 exemplos da classe 'Bad', 217 exemplos da classe 'Good' e 1319 exemplos da classe 'Mid'. Isso faz com que o algoritmo se especialize na classificação de exemplos pertencentes à classe majoritária, desfavorecendo aqueles pertencentes às outras classes. Com o intuito de evitar que isso aconteça, foi realizado o balanceamento da base de forma que os números de exemplos pertencentes à cada classe sejam mais próximos, e para tal, foram utilizadas duas estratégias distintas: *Undersampling* e *Oversampling*.

Utilizando *undersampling*, o conjunto de dados foi subamostrado, de forma que todas as classes possuam quantidades de exemplos iguais à quantidade de exemplos da classe minoritária, ou seja, 63. Para isso, foram selecionadas exatamente 63 amostras aleatórias das classes 'Good' e 'Mid', descartando o restante.

Utilizando *oversampling*, o conjunto de dados foi sobreamostrado, de forma que todas as classes 'Bad' e 'Good' possuam quantidades semelhantes de exemplos. Para isso, foram criados novos exemplos pertencentes a essas classes utilizando a técnica SMOTE (*Synthetic Minority Over-sampling Technique*).

2.3. Treinamentos e Testes Realizados

2.3.1. Arquitetura da Rede

Inicialmente, foi realizado um treinamento utilizando a base de dados completa, com os dados normalizados e sem nenhum tipo de balanceamento, com 70% do conjunto de dados reservado para treinamento e 30% reservado para testes. Nesta etapa, será calculada a acurácia total e a acurácia média por classe do classificador, sem utilizar nenhum método de validação cruzada.

Utilizando essas configurações, foram realizados testes para diferentes arquiteturas de rede, variando a quantidade de camadas internas e também o número de neurônios presentes em cada uma delas. A tabela 1 mostra as arquiteturas utilizadas, bem como as acurácias obtidas para o conjunto de teste em cada uma delas.

Neurônios na 1ª Camada	Neurônios na 2ª Camada	Acurácia total (%)	Acurácia média por classe (%)
5	-	82,92	39,93
11	-	83,54	44,47
22	-	82,29	46,92
5	5	83,96	45,88
5	11	84,58	46,18
5	22	85,21	50,29
11	5	84,58	47,80
11	11	85,00	50,59
11	22	83,75	45,80
22	5	83,12	45,97
22	11	82,71	48,72
22	22	85,21	49,81

Tabela 1 - Arquiteturas utilizadas na classificação e resultados obtidos.

Uma vez que não foram impostas penalidades diferentes para erros cometidos em determinadas classes, vamos considerar como melhor arquitetura aquela que fornece maior valor para a acurácia total, embora isso possa não refletir em uma acurácia por classe elevada. Em etapas posteriores, iremos tratar melhor o balanceamento entre as classes a fim de que o aprendizado da rede seja adequado para todas elas.

A partir deste ponto, a cada teste realizado, a melhor configuração obtida será utilizada para os testes posteriores. Observando os valores obtidos na tabela acima, duas configurações apresentaram maior acurácia geral: **duas camadas com 5 e 11 neurônios** ou 22 neurônios cada. Como a primeira configuração possui uma acurácia por classe média mais elevada, além de possibilitar um treinamento mais rápido, vamos utilizá-la nas próximas etapas.

2.3.2. Número de Épocas

A seguir, foi analisada a influência do número de épocas utilizadas durante o treinamento na acurácia obtida no conjunto de testes. A quantidade de épocas utilizada para cada teste, bem como os resultados obtidos, podem ser observados na tabela 2.

Nº de Épocas	Conjunto de Treinamento		Conjunto de Teste	
	Acurácia Total (%)	Acurácia Média por Classe (%)	Acurácia Total (%)	Acurácia Média por Classe (%)
200	85,61	49,93	83,33	42,67
400	87,58	57,44	84,79	51,27
800	88,03	64,92	82,92	53,56
1000	88,47	64,94	84,58	55,09
2000	85,97	60,98	81,67	51,30

Tabela 2 - Número de épocas utilizadas na classificação e resultados obtidos.

Comparando os valores de acurácia total obtidos para o conjunto de teste, observa-se que o treinamento com 400 epochs foi o que obteve o melhor resultado. Repare que embora o treinamento com um número maior de epochs eleve a acurácia no conjunto de treinamento, à partir de 800 epochs há uma diminuição no score para o conjunto de teste, ilustrando a ocorrência de *overfitting*.

2.3.3. Taxa de Aprendizado e Momentum

Nesta etapa foi analisado como a taxa de aprendizado e o momentum interferem no processo de aprendizado. A tabela 3 contém os valores utilizados para ambos os parâmetros em cada teste realizado, bem como os resultados obtidos para cada um deles.

Taxa de Aprendizado	Momentum	Acurácia Total	Acurácia Média por Classe
0,3	0,3	83,12	40,83
0,3	0,5	82,92	53,99
0,3	0,8	84,58	53,76
0,5	0,3	84,17	46,82
0,5	0,5	84,79	53,03
0,5	0,8	84,79	44,93
0,8	0,3	84,58	45,75
0,8	0,5	84,58	51,66
0,8	0,8	85,42	49,85
1,0	1,0	84,38	45,66

Tabela 3 - Taxa de aprendizagem e Momentum utilizadas na classificação e resultados obtidos.

Observando como a acurácia total varia aumentando-se os valores da taxa de aprendizado e momento, percebe-se uma melhora do resultado ao aproximar os parâmetros de 1. Entretanto, para η e α iguais a 1, a acurácia total foi pior que para o caso em que ambos parâmetros valem 0,8.

2.3.4. Proporções nos Conjuntos de Treinamento e Teste

Por fim, foram variados os tamanhos dos conjuntos de treinamento e de teste. Novamente, foram utilizados 70% dos dados para treinamento, aumentando gradativamente esse valor até atingir a taxa de 95%. Nesta etapa ainda não foi utilizado nenhum método de validação cruzada. A tabela 4 mostra as proporções da base de dados utilizadas para treinamento, e os resultados obtidos para os testes realizados com cada uma delas.

Proporção para Treinamento (%)	Acurácia Total (%)	Acurácia Média por Classe (%)
70	84,17	53,21
75	84,00	42,20
80	80,62	42,31
85	85,00	46,13
90	86,25	47,47
95	83,75	43,94

Tabela 4 - Tamanhos dos conjuntos de treinamento utilizados na classificação e resultados obtidos.

A partir das classificações realizadas, é possível observar que utilizando 70% dos dados para treinamento e 30% dos dados para teste, obtivemos a melhor acurácia média por classe, enquanto que utilizando 90% dos dados para treinamento e 10% para teste, foi obtida a melhor acurácia geral

2.4. Undersampling e Oversampling da Base Original

Utilizando as bases balanceadas conforme descrito na seção 2.2.2, treinamos novos modelos e os avaliamos utilizando *Cross Validation* estratificado com 10 *folds*.

Além disso, após a etapa de validação, treinamos novos modelos utilizando as bases subamostrada e sobreamostrada e testamos sua acurácia na base original. Os resultados encontram-se na Tabela 5 abaixo.

Dataset	Avaliação com Cross Validation		Avaliação no Dataset original	
	Acurácia Total (%)	Acurácia Média por Classe (%)	Acurácia Total (%)	Acurácia Média por Classe (%)
<i>Undersampling</i>	69,84	69,84	52,16	70,00
<i>Oversampling</i>	73,87	51,02	84,37	62,31
Original	80,55	43,39	-	-

Tabela 5 - Acurácia dos modelos treinados utilizando diferentes formatos da base de dados.

Observando os resultados acima, podemos concluir que não há uma abordagem que sempre garanta os melhores resultados. A base de dados que iremos utilizar para construir o modelo irá depender, em grande parte, do objetivo final deste.

Caso queiramos construir um modelo para classificar corretamente vinhos entre as classes 'Good' e 'Bad', a abordagem adotada com o *undersampling* fornecerá o melhor resultado.

Entretanto, caso queiramos identificar com uma acurácia elevada vinhos de qualidade intermediária, treinar um modelo com a base de dados original fornecerá um resultado mais satisfatório.

Além disso, é interessante notar que o classificador não consegue generalizar o conhecimento adquirido durante o treinamento com a base

subamostrada para a base original. Isso se deve ao fato de que não existem exemplos suficientes na base subamostrada para que o classificador consiga identificar satisfatoriamente os outros exemplos presentes na base original, que são majoritariamente da classe 'Mid'. Por outro lado, a acurácia obtida na classificação dos exemplos da classe 'Bad' e 'Good' aumentaram significativamente nesse cenário, pois o treinamento com a base balanceada favoreceu as classes minoritárias.

3. Regressão

3.1. Descrição da Base de Dados

Para o problema de Regressão, foi utilizada uma base de dados contendo 1059 faixas musicais de 33 países/regiões diferentes. Os dados estão armazenados em forma de tabela em um arquivo de texto, onde cada linha corresponde a uma faixa e as primeiras 68 colunas representam características de áudio extraídas de cada uma delas; as duas últimas colunas indicam a latitude e longitude, respectivamente, do ponto considerado como origem da faixa em questão.

3.2. Pré-Processamento dos Dados

3.2.1. Normalização

Assim como foi feito na classificação, os dados foram normalizados de forma que todos os atributos e valores de saída estejam contidos no intervalo (0,1], evitando a saturação dos neurônios da rede. Além disso, como a função de ativação utilizada foi a sigmoide, os possíveis valores de saída da rede também variam de 0 a 1.

3.3. Treinamentos e Testes Realizados

3.3.1. Arquitetura da Rede

Nesta etapa, assim como foi feito para a classificação, serão consideradas diferentes arquiteturas de rede, com uma e duas camadas intermediárias. Entretanto, ao variar o número de neurônios, estes não serão igualados ao número de atributos, à sua metade e ao seu dobro, uma vez que o treinamento para uma camada com apenas 10 neurônios já é lento. Foram consideradas camadas com números de neurônios iguais à 1/4, 1/2 e igual ao número de atributos.

O treinamento foi realizado sobre a base de dados completa com os dados normalizados, utilizando 70% do conjunto de dados para treinamento e apenas 100 épocas. Nesta etapa, como medida de desempenho, foi calculado o erro quadrático médio entre a saída desejada e a saída obtida do regressor, e não foi utilizado nenhum método de validação cruzada.

A tabela 6 mostra as arquiteturas utilizadas, bem como o resultado obtido para cada uma delas.

Neurônios na 1ª Camada	Neurônios na 2ª Camada	Erro quadrático Médio (Teste)
17	-	0,0373
34	-	0,0316
68	-	0,0332
17	17	0,0412
17	34	0,0344
17	68	0,0369
34	17	0,0419
34	34	0,0401
34	68	0,0374
68	17	0,0358
68	34	0,0390
68	68	0,0417

Tabela 6 - Arquiteturas utilizadas na regressão e resultados obtidos.

3.3.2. Número de Épocas

A seguir, foi analisada a influência do número de épocas utilizadas durante o treinamento nos resultados obtidos para o conjunto de testes. A quantidade de

épocas utilizada para cada teste, bem como os resultados obtidos, podem ser observados na tabela 7.

Nº de Épocas	Erro Quadrático Médio (Treinamento)	Erro quadrático Médio (Teste)
100	0,0238	0,0316
200	0,0199	0,0347
400	0,0110	0,0397
800	0,0073	0,0453
1000	0,0054	0,0486

Tabela 7 - Números de épocas utilizados na regressão e resultados obtidos.

Analizando os dados acima, pode-se observar que o erro quadrático médio sempre diminui para o conjunto de treinamento ao aumentar a epoch. Entretanto, analisando os valores obtidos para o conjunto de teste, observa-se que o erro já começa a subir à partir de 200 epochs, demonstrando a perda da capacidade de generalização da rede (*overfitting*). Portanto, utilizaremos 100 epochs como o número de iterações para as etapas seguintes.

3.3.3. Taxa de Aprendizado e Momentum

Nesta etapa foi analisado como a taxa de aprendizado e o momentum interferem no processo de aprendizado. A tabela 8 contém os valores utilizados para ambos os parâmetros em cada teste realizado, bem como o erro quadrático médio obtido para os testes realizados com cada um deles.

Taxa de Aprendizado	Momentum	Erro Quadrático Médio
0,3	0,3	0,0336
0,3	0,5	0,0309
0,3	0,8	0,0322
0,5	0,3	0,0315
0,5	0,5	0,0332
0,5	0,8	0,0396
0,8	0,3	0,0325
0,8	0,5	0,0344
0,8	0,8	0,0505
1,0	1,0	0,0356

Tabela 8 - Taxa de aprendizagem e Momentum utilizados na regressão e resultados obtidos.

A partir dos dados da tabela, é possível observar que, contrariamente à classificação, um aumento nos valores dos parâmetros de aprendizado e momentum tende a aumentar o erro quadrático médio para o mesmo número de epochs. Assim como para a classificação, quando *eta* e *alpha* assumem o valor 1, há uma quebra nessa tendência.

3.3.4. Proporção nos Conjuntos de Treinamento e Teste

Também foram variados os tamanhos dos conjuntos de treinamento e de teste. Novamente, foram utilizados 70% dos dados para treinamento, aumentando gradativamente esse valor até atingir a taxa de 95%. Nesta etapa ainda não foi utilizado nenhum método de validação cruzada. A tabela 9 mostra as proporções da base de dados utilizadas para treinamento, e os resultados obtidos para os testes realizados com cada uma delas.

Proporção para Treinamento (%)	Erro Quadrático Médio
70	0,0320
75	0,0312
80	0,0326
85	0,0392
90	0,0272
95	0,0399

Tabela 9 - Tamanhos dos conjuntos de treinamento utilizados na regressão e resultados obtidos

Observa-se que o menor erro quadrático médio obtido foi de 0,0272, utilizando 90% dos dados disponíveis na base para treinamento e apenas 10% deles para teste.

3.4. Avaliação do Modelo Final

Após encontrar a melhor configuração possível para o modelo, foi realizada uma avaliação final utilizando o método de validação cruzada *Stratified K-Fold*, com 10 *folds*, e foi calculada a média aritmética dos erros quadráticos médios (MSE) obtidos para cada iteração. Os resultados obtidos utilizando essa estratégia, para os conjuntos de treinamento e de teste, podem ser observados na tabela 10 abaixo.

MSE Médio	
Conjunto de Treinamento	Conjunto de Teste
0,0248	0,0339

Tabela 10 - MSE Médios obtidos no Conjunto de Treinamento e de Teste, utilizando o método de validação cruzada *Stratified K-Fold*.