Predictive Modeling of Obesity Risk Using Linear Regression, PCR and Neural Networks

Juan de Souza Holanda
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
juanholanda@alu.ufc.br

Matheus de Castro Vieira
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
matheusdcastro@alu.ufc.br

Diego Duarte de Lima
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
diegolimaufc@alu.ufc.br

Mateus Andrade Maia
Federal University of Ceará (UFC)
Department of Teleinformatics (DETI)
Fortaleza, Brazil
mateusmaia45@alu.ufc.br

***Abstract*—A modelagem estatística de dados biomédicos desempenha um papel fundamental na predição de riscos à saúde e no suporte à tomada de decisão clínica. Neste trabalho, são avaliados diferentes modelos de regressão aplicados ao conjunto de dados "Obesity or CVD Risk", com o objetivo de estimar o nível de obesidade de indivíduos a partir de atributos demográficos, físicos e comportamentais. Foram investigados os modelos de Regressão Linear Ordinária (OLS), Regressão Ridge (L2), Regressão por Componentes Principais (PCR) e Redes Neurais Artificiais do tipo Multilayer Perceptron (MLP). Os dados foram previamente submetidos a etapas de codificação de variáveis categóricas, transformação de assimetria e padronização. A avaliação dos modelos foi realizada por meio de validação cruzada e métricas de desempenho no conjunto de teste, utilizando o Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2). Os resultados demonstram que os modelos lineares apresentaram excelente desempenho preditivo, com valores de R^2 superiores a 0,95, enquanto a rede neural apresentou desempenho comparável, confirmando a forte relação linear entre os atributos antropométricos e o nível de obesidade.**

***Index Terms*—Regressão Linear, Principal Component Regression, Redes Neurais, Predição de Obesidade**

I. Introdução

A obesidade é reconhecida como um dos principais problemas de saúde pública da atualidade, estando diretamente associada ao desenvolvimento de doenças crônicas como diabetes, hipertensão arterial e enfermidades cardiovasculares. O crescimento acelerado dessa condição em diferentes faixas etárias tem motivado o uso de técnicas computacionais para análise e modelagem de dados clínicos, com o objetivo de apoiar estratégias de prevenção, diagnóstico e intervenção em saúde.

Nesse contexto, os métodos de regressão assumem papel central na construção de modelos preditivos capazes de estimar variáveis de interesse a partir de múltiplos fatores explicativos. A Regressão Linear Ordinária (OLS) é amplamente utilizada por sua simplicidade e interpretabilidade, porém pode apresentar limitações em cenários de multicolinearidade entre os preditores. Para contornar essas limitações, técnicas como a Regressão Ridge (regularização L2) e a Regressão por Componentes Principais (PCR) têm sido amplamente empregadas na literatura, pois introduzem mecanismos de controle da variância e redução de dimensionalidade, respectivamente. Além disso, modelos de aprendizado de máquina baseados em Redes Neurais Artificiais vêm ganhando destaque por sua capacidade de modelar relações não lineares complexas.

Diversos trabalhos têm explorado a aplicação de modelos de regressão na área médica, incluindo previsão de risco cardiovascular, análise de composição corporal e diagnóstico assistido por computador [2]–[4]. Em particular, bases de dados relacionadas à obesidade permitem investigar de forma quantitativa a influência de fatores como idade, peso, altura, hábitos alimentares e nível de atividade física sobre os diferentes graus dessa condição clínica.

Diante desse cenário, o presente trabalho tem como objetivo realizar um estudo comparativo entre diferentes modelos de regressão aplicados ao conjunto de dados "Obesity or CVD Risk", composto por 2111 amostras e 17 atributos. São avaliados os desempenhos dos modelos OLS, Ridge, PCR e uma Rede Neural do tipo MLP, considerando métricas de erro no conjunto de teste e validação cruzada. A partir dessa análise, busca-se identificar o modelo mais adequado para a predição do nível de obesidade, bem como compreender o impacto das técnicas de regularização, redução de dimensionalidade e modelagem não linear sobre a capacidade preditiva dos modelos.

II. Metodologia

A. Descrição do Conjunto de Dados

O conjunto de dados utilizado neste artigo consiste em 2111 observações e 17 colunas considerando a variável alvo. O objetivo é prever o nível de obesidade (NObeyesdad) com base em 16 atributos, que incluem dados demográficos (Idade, Gênero), características físicas (Altura, Peso) e hábitos alimentares/comportamentais (e.g., consumo calórico, atividade

física, uso de álcool). A variável alvo foi codificada ordinalmente, variando de 0 (Peso Insuficiente) a 6 (Obesidade Tipo III), permitindo uma abordagem de regressão para estimar o grau de risco.

1) Estrutura e características: O conjunto de dados é composto por variáveis preditores agrupadas em três categorias principais: atributos demográficos, hábitos alimentares e condição física. A variável resposta (alvo) representa o nível de obesidade estimado para cada indivíduo.

B. Pré-processamento e Análise dos dados

Sabemos que a qualidade dos dados é determinante para o desempenho de modelos de aprendizagem de máquina. O conjunto de dados original continha uma mistura de atributos numéricos (Age, Height, Weight) e categóricos (CAEC, MTRANS). A fim de melhorar o desempenho do modelo, foi implementado uma pipeline de pré-processamento em três etapas principais: codificação de variáveis, tratamento de assimetria e padronização.

1) Codificação de Variáveis Categóricas: : Para converter os dados textuais em representações numéricas adequadas para operações algébricas, foi adotadp estratégias distintas baseadas na natureza semântica de cada atributo:

- Mapeamento Ordinal (Manual): Variáveis que possuem uma hierarquia intrínseca foram mapeadas para inteiros preservando sua ordem. Para os atributos $CAEC$ (consumo de comida entre refeições) e $CALC$ (consumo de álcool), utilizou-se a escala: 'no': 0, 'Sometimes': 1, 'Frequently': 2, 'Always': 3. A variável alvo NObeyesdad também foi codificada ordinalmente de 0 a 6, permitindo tratar a classificação de obesidade como um problema de regressão. Permitindo que o modelo "compreenda" que o aumenta da escala numérica influencia diretamente a variável alvo
- Label Encoding (Binário): Aplicado a variáveis booleanas como $Gender$, $family_history_with_overweight$, $FAVC$, $SMOKE$ e SCC, convertendo-as para 0 ou 1
- One-Hot Encoding: Aplicado à variável nominal $MTRANS$ (meio de transporte). Como não existe ordem matemática entre "Carro" e "Transporte Público", criaram-se colunas dummy ($MTRANS_Automobile$, $MTRANS_Walking$) para evitar que o modelo interpretasse falsas hierarquias.

2) Análise de Correlação: A análise das variáveis numéricas revelou assimetrias em alguns preditores. A variável Age, por exemplo, apresenta uma distribuição enviesada, indicando uma predominância de indivíduos mais jovens e outliers em idades avançadas. Para mitigar o impacto desses extremos e aproximar a distribuição de uma normal — uma premissa desejável para a estabilidade dos estimadores de Mínimos Quadrados Ordinários (OLS) — aplicou-se uma transformação logarítmica aos dados de idade.

3) Padronização e Viés: Para garantir que todas as variáveis contribuíssem equitativamente para o gradiente de erro e para a estabilidade numérica da inversão de matrizes, aplicou-se também uma padronização dos dados. Os dados foram transformados tal que μ=0 e σ=1, ou seja, agrupados para uma média entre 0 e 1.

Por fim, para incorporar o intercepto (β_0) na equação vetorial do modelo linear ($y = X\beta + \epsilon$), adicionou-se uma coluna de uns (1s) à matriz de características X. O conjunto final foi dividido em 80% para treinamento e 20% para teste, garantindo a avaliação em dados não vistos.

C. Implementação do modelo de regressão linear (OLS)

Para quantificar a relação entre os atributos físico-comportamentais e o nível de obesidade, adotou-se o modelo de Regressão Linear Múltipla. O objetivo do método de Mínimos Quadrados Ordinários (OLS) é estimar o vetor de coeficientes β que minimiza a Soma dos Erros Quadráticos (SSE) entre os valores observados y e os valores preditos ŷ

A implementação foi realizada utilizando a biblioteca NumPy para manipulação algébrica, seguindo a formulação matricial. O modelo é definido por:

$$y = X\beta + \epsilon$$

A relevância desta abordagem para o problema em questão reside na interpretabilidade direta dos coeficientes β. Cada componente β_j quantifica a mudança esperada no nível de obesidade para uma unidade de variação no preditor j, mantendo-se constantes as demais variáveis. Isso permite não apenas prever o risco, mas isolar o impacto individual de fatores como o consumo de álcool ou o uso de transporte público

O problema de aprendizagem consiste em encontrar o vetor de parâmetros $\hat{\beta}$ que minimize a Função de Custo $J(\beta)$, definida como a Soma dos Quadrados dos Resíduos (RSS):

$$J(\beta) = ||y - X\beta|| = \sum_{i=1}^{n}(y_i - \hat{y_i})$$

Diferentemente de métodos iterativos, a abordagem OLS permite uma solução analítica fechada. Ao derivar a função de custo em relação a β e igualar o gradiente a zero ($\nabla\beta J(\beta) = 0$), obtém-se a Equação Normal:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

O modelo resultante oferece, portanto, o Melhor Estimador Linear Não-Viesado, conforme o Teorema de Gauss-Markov, assegurando que as estimativas de risco de obesidade sejam as mais precisas possíveis dentro da classe de modelos lineares.

D. Regressão Linear Penalizada (Ridge/L2)

Embora o método de Mínimos Quadrados Ordinários (OLS) seja eficiente, ele pode apresentar alta variância e risco de overfitting em cenários de alta dimensionalidade ou multicolinearidade. Para investigar a robustez do modelo, implementou-se a Regressão Ridge (Penalização L2).Diferente do OLS, a Regressão Ridge adiciona um termo de regularização à função de custo, penalizando a magnitude dos coeficientes vetoriais β. O objetivo passa a ser minimizar:

$$J(\beta) = MSE(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$$

Onde $\lambda \geq 0$ é o hiperparâmetro de complexidade.

- A implementação foi realizada manualmente (from scratch) utilizando o algoritmo de Gradiente Descendente, cuja regra de atualização dos pesos é dada por: $\beta_{new} = \beta_{old} - \alpha(\nabla MSE + \lambda\beta_{old})$. O intercepto ($\beta_0$) não foi penalizado.
- A escolha do λ ótimo foi determinada via Validação Cruzada k-fold (com $k = 5$) no conjunto de treinamento, explorando um espaço de busca $\lambda \in [0, 100]$.
- Para validar a corretude do algoritmo manual, os resultados finais foram comparados com a implementação de referência da biblioteca Scikit-Learn.

Optou-se pela implementação da penalização L2 (Ridge) em detrimento da L1 (Lasso) por duas razões principais: primeiramente, a diferenciabilidade da função de custo L2 em todos os pontos facilita a implementação estável do algoritmo de Gradiente Descendente; em segundo lugar, a penalização L2 é particularmente eficaz no tratamento da multicolinearidade entre variáveis antropométricas, preservando a contribuição de todos os preditores no modelo final.

E. Regressão por Componentes Principais (PCR)

A Regressão por Componentes Principais (Principal Component Regression – PCR) foi empregada como uma estratégia para reduzir a dimensionalidade do problema e mitigar possíveis efeitos de multicolinearidade entre os preditores. O método consiste na aplicação inicial da Análise de Componentes Principais (PCA) sobre a matriz de características padronizada, seguida do ajuste de um modelo de Regressão Linear Ordinária (OLS) sobre os componentes selecionados.

Inicialmente, os dados foram centralizados por meio da subtração da média de cada variável. Em seguida, a decomposição em valores singulares (Singular Value Decomposition – SVD) foi aplicada à matriz centralizada, permitindo a obtenção dos autovetores associados aos maiores autovalores. Esses autovetores definem as direções de máxima variância dos dados. Os preditores originais foram então projetados nesse subespaço reduzido, gerando os chamados *scores* principais.

A regressão linear foi ajustada sobre esses *scores*, e os coeficientes finais foram reconstruídos no espaço original das variáveis, resultando no modelo PCR final. A escolha do número ótimo de componentes principais foi realizada por meio de validação cruzada k-fold com $k = 10$, utilizando exclusivamente o conjunto de treinamento. Para cada valor de componentes, o erro quadrático médio (RMSE) foi calculado, sendo selecionado o número de componentes que minimizou o erro médio de validação.

Para validar a implementação manual do PCR, os resultados obtidos foram posteriormente comparados com a implementação de referência da biblioteca *Scikit-Learn*, utilizando as classes `PCA` e `LinearRegression`, adotando-se exatamente o mesmo número de componentes selecionado no processo de validação cruzada.

F. Rede Neural para Regressão Linear

Diferentemente dos modelos lineares usados anteriormente, as *Neural Networks* (NN) aprendem a forma da função que relaciona os dados de entrada e saída sozinhas, sem supor uma relação linear entre elas. A vantagem de usar modelos de rede neurais em problemas reside na sua capacidade de modelar relações não lineares mais complexas. Embora, para isso, exijam mais recursos computacionais e sejam mais sensíveis ao *overfitting*.

Existem diferentes arquiteturas de redes neurais, cada uma mais adequada a diferentes situações. Para o dataset usado - aproximadamente 2 mil amostras - escolher um modelo com redes profundas seria arriscado devido a grande possibilidade de haver *overfitting*. Então, a arquitetura *Multilayer FeedForward Network*, também conhecida como *Multilayer Perceptron* (MLP) foi escolhida, com duas camadas densamente conectadas - uma com 64 e outra com 32 neurônios.

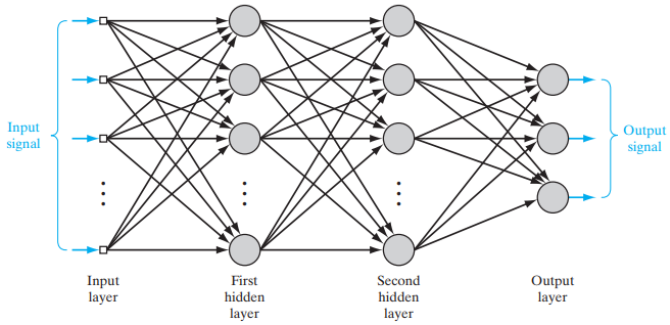

Fig. 1. Generalização de uma MLP com dois *hidden layers*

Dessa forma, para um problema de regressão, a camada de saída deve possuir apenas um neurônio (sem ativação), produzindo um valor escalar contínuo.

A função de ativação usada nas camadas ocultas foi a *Rectified Linear Unit*(ReLU), definida como:

$$ReLU(x) = max(0, x)$$

A ReLU introduz não linearidade na rede, permitindo que a MLP modele relações mais complexas entre os preditores. Se todas as camadas fossem apenas combinações lineares, a rede neural não teria melhor performance do que uma regressão linear comum.

Os cálculos dentro de uma MLP são baseados em operações de combinação linear seguidos de funções de ativação. Para

cada neurônio j em uma camada, o valor de ativação antes da função não linear é:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n)$$

onde $y_i(n)$ é a saída do neurônio i da camada anterior, e w_{ij} é o peso sináptico.

Depois de calculado $v_j(n)$, a função de ativação é aplicada e a saída um neurônio j é:

$$y_i(n) = \varphi_j(v_j(n))$$

Durante o treinamento de uma rede neural, o conjunto de treino é exposto ao modelo diversas vezes. Cada passagem completa por todo o conjunto de treino é chamada de época (*epoch*). Em cada época, o algoritmo de otimização ajusta os pesos da rede com base no gradiente do erro entre a predição e o valor real. Os testes mostraram que para número de épocas muito maiores que 40, o resultado é pouco expressivo em comparação com o aumento significativo do esforço computacional. Já para valores menores, *underfitting* passa a ser um risco.

III. Resultados e Discussão

A. Modelo de Regressão Linear (OLS)

Para validar a robustez matemática da implementação manual, os resultados foram comparados diretamente com funções conhecidas. A avaliação foi realizada no conjunto de teste (20% das amostras), utilizando as métricas de Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2). Os resultados comparativos são apresentados na Tabela [N].

Implementação	**RMSE (Erro Médio)**	**R^2 (Coef. Det.)**
Manual (Scratch)	0.41883	0.95588
Scikit-Learn	0.41883	0.95588

TABLE I
DESEMPENHO NO CONJUNTO DE TESTE (OLS)

Observa-se que a implementação matemática "manual" alcançou uma convergência exata com a biblioteca padrão até a quinta casa decimal, validando o cálculo algébrico dos coeficientes β. O valor de $R^2 \approx 0.96$ indica que o modelo linear simples foi capaz de explicar aproximadamente 95,6% da variabilidade dos dados de obesidade no conjunto de teste. Este alto desempenho sugere que, após o pré-processamento adequado, a relação entre os atributos físico-comportamentais e o nível de obesidade é predominantemente linear.

A Fig. 2 ilustra a dispersão entre os valores reais (eixo X) e os valores preditos pelo modelo OLS (eixo Y). A forte concentração dos pontos ao longo da diagonal principal confirma a precisão do modelo e a ausência de viés significativo (underfitting ou overfitting) para este conjunto de dados.

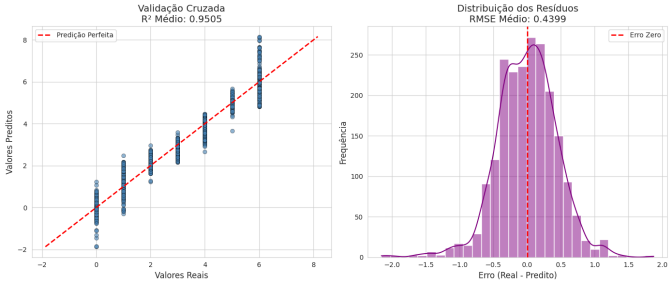

Fig. 2. Comparação entre os valores reais de obesidade e as predições do modelo OLS. A linearidade dos pontos indica alta acurácia na predição.

Além disso, a análise dos resíduos, representada pelo histograma de distribuição de erros, exibe um comportamento Gaussiano (Normal) com média centrada em zero. Essa característica é fundamental para validar as premissas estatísticas da Regressão Linear, demonstrando que o modelo é não-viesado — ou seja, não apresenta tendências sistemáticas de superestimar ou subestimar os valores — e que os desvios das predições são aleatórios e consistentes.

B. Regressão Linear Penalizada (Ridge/L2)

1) Análise de Regularização e Seleção de Hiperparâmetros: A etapa de validação cruzada (*5-fold*) para a Regressão Ridge teve como objetivo identificar se a introdução de viés (penalidade L2) reduziria o erro de generalização. Ao testar o espaço de busca para o hiperparâmetro de regularização, observou-se que o menor Erro Quadrático Médio (RMSE) foi obtido com $\lambda = 0$.

Conforme ilustrado na Fig. 3 (Perfil de Validação Cruzada), o comportamento do erro é monotônico crescente: à medida que a penalidade λ aumenta, o RMSE (linha vermelha) sobe e o R^2 (linha azul) diminui.

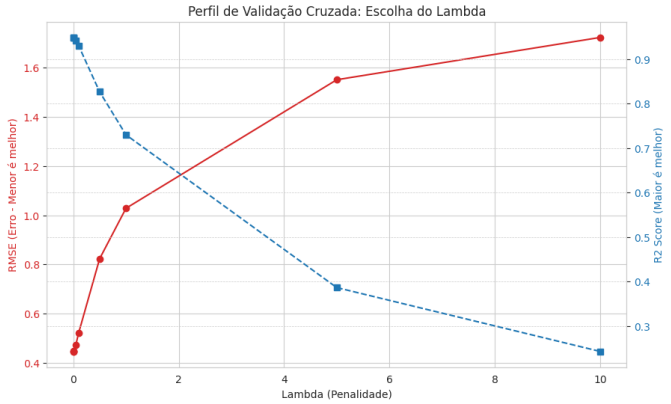

Fig. 3. Perfil de Validação Cruzada (k=5). Observa-se que o menor erro ocorre em $\lambda = 0$, indicando ausência de overfitting significativo.

Este resultado sugere que o modelo linear não sofria de *overfitting* no conjunto de dados original. A explicação para este fenômeno reside na análise de correlação (Fig. 4), que evidencia uma relação linear fortíssima ($r > 0.9$) entre a variável *Weight* e a variável alvo *NObeyesdad*. Dada a natureza clínica do problema (onde a obesidade é uma função direta do peso e altura) e a razão favorável entre o número de

amostras ($N = 2111$) e preditores, o estimador OLS simples demonstrou ser a solução mais robusta e estável, tornando a regularização desnecessária.

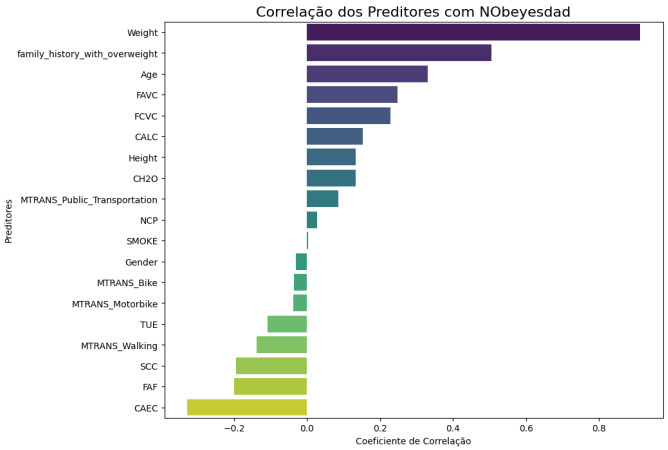

Fig. 4. Correlação dos preditores com a variável alvo. A predominância da variável peso (Weight) explica a estabilidade do modelo linear.

2) Desempenho do Modelo e Validação Comparativa: O modelo final, treinado com os parâmetros ótimos ($\lambda = 0$), foi avaliado no conjunto de teste (20% dos dados, não vistos no treinamento). Para garantir a integridade da solução desenvolvida manualmente (*Scratch*), comparou-se seu desempenho com a biblioteca padrão (*Scikit-Learn*). A Tabela II apresenta os resultados obtidos.

Implementação	**RMSE (Erro Médio)**	R^2 **(Coef. Det.)**
Manual (Scratch)	0.41947	0.95575
Scikit-Learn	0.41883	0.0.95589

TABLE II
DESEMPENHO NO CONJUNTO DE TESTE (RIDGE REGRESSION)

A convergência dos resultados até a quarta casa decimal valida a implementação do algoritmo de Gradiente Descendente. O coeficiente de determinação $R^2 \approx 0.96$ indica que o modelo é capaz de explicar 95,6% da variância dos dados.

A Fig. 5 apresenta a dispersão entre os valores reais (eixo X) e os valores preditos (eixo Y). A sobreposição dos pontos das duas implementações e o alinhamento com a diagonal ideal confirmam a ausência de viés sistemático nas predições.

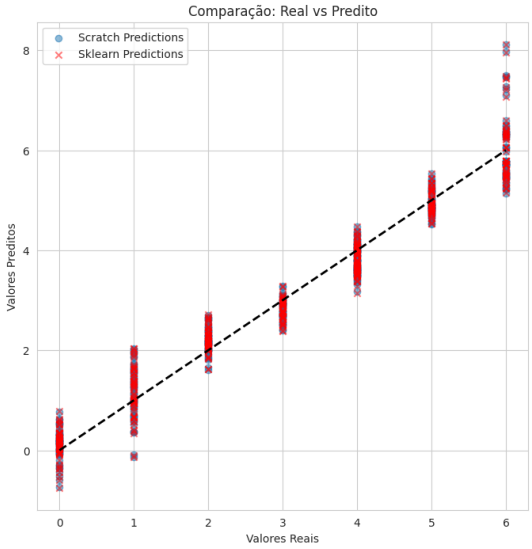

Fig. 5. Comparação entre valores reais e preditos no conjunto de teste. A sobreposição das séries valida a implementação manual.

Em síntese, a análise demonstra que, para este conjunto de dados, uma abordagem de regressão linear bem parametrizada (com tratamento logarítmico da idade e padronização) é suficiente para obter alta acurácia preditiva, superando a necessidade de penalização L2 complexa.

C. Regressão por Componentes Principais (PCR)

O modelo PCR, treinado com o número ótimo de componentes determinado por validação cruzada, foi avaliado no conjunto de teste (20% das amostras), utilizando-se as métricas de Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). Para garantir a corretude da implementação manual desenvolvida, os resultados foram comparados diretamente com a implementação de referência da biblioteca *Scikit-Learn*. A Tabela III apresenta os resultados obtidos.

Implementação	**RMSE (Erro Médio)**	R^2 **(Coef. Det.)**
Manual (Scratch)	0.41883	0.95588
Scikit-Learn	0.41883	0.95588

TABLE III
DESEMPENHO DO MODELO PCR NO CONJUNTO DE TESTE.

Observa-se que a implementação manual do PCR apresentou concordância numérica exata com a biblioteca padrão até a quinta casa decimal, validando integralmente a formulação matemática empregada. O valor de $R^2 \approx 0.956$ indica que aproximadamente 95,6% da variabilidade da variável alvo é explicada pelo modelo, evidenciando excelente capacidade de generalização.

A Fig. 6 ilustra a dispersão entre os valores reais (eixo X) e os valores preditos pelo modelo PCR (eixo Y). Observa-se forte concentração dos pontos ao longo da diagonal principal, caracterizando elevada acurácia nas predições e ausência de viés sistemático relevante. A completa sobreposição entre as predições do modelo manual e do modelo da biblioteca

confirma a equivalência algébrica e numérica entre as duas abordagens.

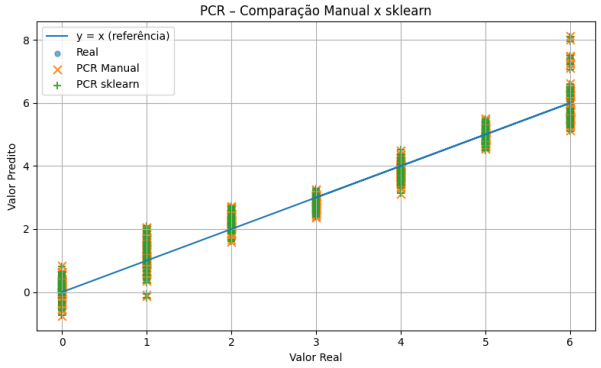

Fig. 6. Comparação entre os valores reais e os valores preditos pelo modelo PCR no conjunto de teste.

Em síntese, os resultados demonstram que a aplicação da redução de dimensionalidade via PCA, seguida de regressão linear, produziu um modelo altamente estável e preciso. O desempenho do PCR mostrou-se superior e comparável aos melhores modelos lineares avaliados neste estudo.

D. Modelo de Rede Neural para Regressão Linear

O modelo de MLP, treinado com 40 épocas, usando MSE como função de perda tem seu desempenho mostrado na figura a seguir

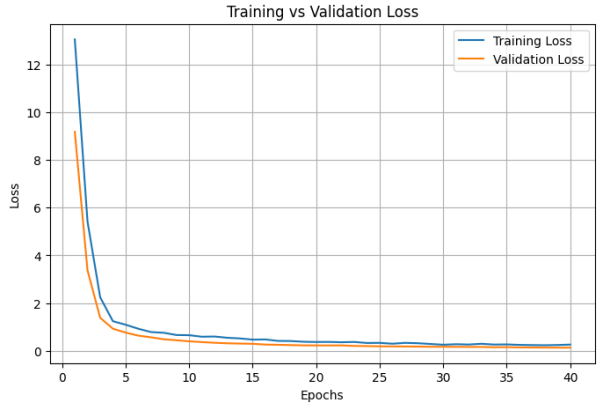

Fig. 7. Perdas no Treinamento e Perdas na Validação ao Longo das Épocas

Pelo gráfico, é nítido que tanto a perda no treinamento quanto na validação decrescem e rapidamente convergem. Esse comportamento ótimo indica que o modelo é capaz de aprender padrões reais e generalizar os conhecimentos de dados com os quais jé teve contato para dados aos quais o modelo nunca tinha sido exposto.

Como comentado anteriormente, testes com valores de épocas mais altos mostraram que não há melhora significativa quando comparada ao aumento de consumo computacional.

Um dos gráficos presentes no **Colab** também permite interpretar que para os valores preditos pelo modelo estão, em geral, bem próximos aos valores reais. No entanto, também é destacável que, para os valores 1, 2 e 6 - classes `Normal_Weight`, `Overweight_Level_I` e `Obesity_Type_III`, respectivamente - há uma maior dispersão dos valores preditos.

TABLE IV
COMPARAÇÃO DE DESEMPENHO ENTRE OS MODELOS

Modelo	**RMSE (Teste)**	**R^2 (Teste)**
Rede Neural (MLP)	0.4049	0.9588
Regressão Linear	0.4220	0.9552

Por fim quando comparados os parâmetros de acurácia entre o modelo MLP e de regressão linear, é perceptível que os valores tanto de R^2 quanto de $RMSE$ são bem próximos. O modelo MLP teve $RMSE$ um pouco menor, mas apresentou R^2 ligeiramente maior do que a regressão linear. Portanto, essa proximidade entre os valores mostra que uma relação linear é capaz de descrever bem o vínculo entre preditores e variável de interesse, embora o desempenho levemente superior da rede neural signifique que há presença de relações não lineares, ainda que bem fraca. Assim, o problema não é altamente não linear, mas também não pode ser limitado a padrões estritamente lineares.

IV. CONCLUSÃO

Neste trabalho, foi realizado um estudo comparativo entre diferentes modelos de regressão aplicados à predição do nível de obesidade a partir do conjunto de dados "Obesity or CVD Risk". Os modelos avaliados — Regressão Linear (OLS), Regressão Ridge, Regressão por Componentes Principais (PCR) e Rede Neural do tipo MLP — apresentaram desempenho elevado no conjunto de teste, com coeficiente de determinação R^2 superior a 0,95.

Os resultados mostraram que a Regressão Ridge não trouxe ganhos relevantes em relação ao modelo OLS, indicando que o problema não apresenta forte sensibilidade à multicolinearidade. De forma semelhante, o PCR apresentou desempenho equivalente ao OLS, demonstrando que a redução de dimensionalidade não comprometeu a capacidade preditiva do modelo.

A Rede Neural apresentou desempenho comparável aos modelos lineares, sugerindo que, para este conjunto de dados, a relação entre os atributos e a variável alvo é predominantemente linear. Assim, conclui-se que os modelos lineares são suficientes, eficientes e adequados para a predição do nível de obesidade no contexto analisado.

REFERENCES

[1] Kaggle, "Obesity or CVD Risk Dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster/data

[2] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, New York, NY: Springer, 2013.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, NY: Springer, 2008.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in Python*, New York, NY: Springer, 2023.

[5] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Upper Saddle River, NJ: Pearson, 2009.