Análise e Predição de Diabetes

Dataset Pima Indians

Professor: Vitor Casadei
Aluno: Douglas F Azevedo
Email: dfa3@cesar.school
Tema: Predição do Diabetes por RNA

Sumário

Imagens em alta resolução

Para visualizar as figuras deste documento em alta resolução, acesse Análise e Predição de Diabetes: figuras

Introdução

Este trabalho apresenta uma análise exploratória e desenvolvimento de um modelo preditivo para diagnóstico de diabetes, utilizando o dataset Pima Indians Diabetes Database.

O conjunto de dados contém informações clínicas de 768 pacientes, todas mulheres acima de 21 anos pertencentes à etnia Pima Indian. O objetivo é classificar se o paciente tem diabetes (1) ou não (0) com base em variáveis clínicas diversas.

Descrição do Dataset

O dataset apresenta 9 colunas, sendo 8 variáveis independentes e 1 variável alvo ("*Outcome*"). As principais variáveis são:

- **Pregnancies:** Número de gestações
- **Glucose:** Concentração plasmática de glicose
- **BloodPressure:** Pressão arterial diastólica (mm Hg)
- **SkinThickness:** Espessura da dobra cutânea do tríceps (mm)
- **Insulin:** Nível de insulina sérica
- **BMI:** Índice de massa corporal $IMC = \frac{massa}{altura^2}$
- **DiabetesPedigreeFunction:** Função de pedigree diabético (probabilidade baseada em histórico familiar)
- **Age:** Idade (anos)
- **Outcome:** Diagnóstico de diabetes (0 = não diabético, 1 = diabético)

Análise Exploratória dos Dados

A paciente mais velha possui 81 anos e a mais nova 21. maioria das pacientes tem entre 21 e 40 anos.

Glicose e BMI apresentam maior concentração em intervalos de risco para diabetes.

A análise inicial mostrou que os dados estão completos, porém apresentam diversos valores zero em colunas como Glicose, Pressão Arterial, Espessura da Pele, Insulina e BMI, que são biologicamente incompatíveis, indicando dados ausentes ou inconsistentes.

RowID	Pregnanc... Number (Inte...	Glucose Number (Inte...	BloodPre... Number (Inte...	SkinThick... Number (Inte...	Insulin Number (Inte...	BMI Number (Float)	Diabetes... Number (Float)	Age Number (Inte...	Outcome Number (Inte...
Row0	6	148	72	35	0	33.6	0.627	50	1
Row1	1	85	66	29	0	26.6	0.351	31	0
Row2	8	183	64	0	0	23.3	0.672	32	1
Row3	1	89	66	23	94	28.1	0.167	21	0
Row4	0	137	40	35	168	43.1	2.288	33	1
Row5	5	116	74	0	0	25.6	0.201	30	0
Row6	3	78	50	32	88	31	0.248	26	1
Row7	10	115	0	0	0	35.3	0.134	29	0
Row8	2	197	70	45	543	30.5	0.158	53	1
Row9	8	125	96	0	0	0	0.232	54	1
Row10	4	110	92	0	0	37.6	0.191	30	0
Row11	10	168	74	0	0	38	0.537	34	1
Row12	10	139	80	0	0	27.1	1.441	57	0

Figura 1: Exemplos de valores zerados no dataset

Após tratamento desses dados, desconsiderando as tuplas que possuem zeros em colunas importantes, o dataset final ficou com 392 registros, nos quais servirão como insumo para o treinamento da RNA.

Correlações

- **Glicose** apresentou correlação moderada positiva (0,51) com o diagnóstico de diabetes, sendo o preditor mais forte.
- **Idade** e Insulina **ficaram** em segundo com a correlação moderada positiva (0,35 e 0,30).
- **Índice de Massa Corporal** (BMI) mostrou correlação moderada (0,27).
- **Espessura de pele** e **Número de gestações** tiveram correlação fraca positiva (0,25 ambas)
- **Ocorrência familiar** teve correlação fraca (0,21)
- **Pressão arterial** demonstrou fraca correlação (0,19)

First column name String	Second column name ↓ String	Correlation value Number (Float)
Pregnancies	Outcome	0.257
Glucose	Outcome	0.516
BloodPressure	Outcome	0.193
SkinThickness	Outcome	0.256
Insulin	Outcome	0.301
BMI	Outcome	0.27
DiabetesPedigreeFunction	Outcome	0.209
Age	Outcome	0.351

Figura 2: Correlações com a coluna "Outcome"

Divisão

Os dados foram divididos da seguinte forma:

- 75% para treino e validação, destes sendo:
  - 85% usados para treinamento do modelo; e
  - 15% para validação, monitorando o desempenho e evitando *overfitting*.
- 25% reservados para o teste final do modelo

Após a divisão, os dados foram normalizados para ajustar as variáveis em escalas compatíveis, o que facilita e otimiza o treinamento da rede.

Arquitetura

Para treinar o modelo preditivo de diabetes foi utilizada uma rede neural artificial multi camada. A camada de entrada foi configurada com o *shape* 8 e o tipo de dado *Float 32*, que representa todas as variáveis com exceção da "*outcome*". As camadas densas foram configuradas com a função de ativação ***ReLU*** $(x) = \max(0, x)$ nas camadas ocultas, com **16**, **32** e **16** neurônios respectivamente e ***Softmax*** $P(y = 0|1) = \frac{e^{x_1}}{e^{x_0}+e^{x_1}}$ na camada de saída para classificação binária.

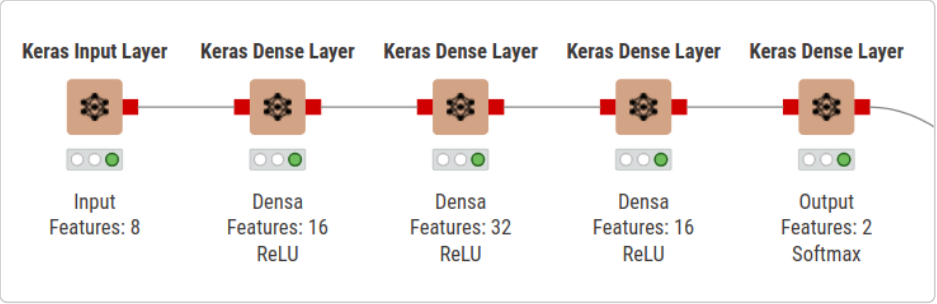

Figura 3: Camadas, neurônios e funções de ativação

Para o nó ***Keras Network Learner***, a função de saída alvo foi configurada como ***Binary Cross Entropy*** $\mathcal{L} = -\left(y \log(\sigma(\hat{y})) + (1 - y) \log(1 - \sigma(\hat{y}))\right)$ para a classificação binária. E o otimizador foi o ***Adam***, que apresentou melhor acurácia para este caso. Configurei o número de épocas para 64 e os dados de ***Training batch size*** e ***Validation batch size*** com **32** cada.

Condições de parada

Apesar de não ter sido parte do escopo da disciplina, configurei também condições de parada em ***Advanced Options***, que foram cruciais para conseguirmos uma acurácia mais definida (pois estavam randomicamente variando a precisão).

- *Monitored quantity: **Training loss (total)***
- *Min. delta: **0.01***
- *Patience: **10***

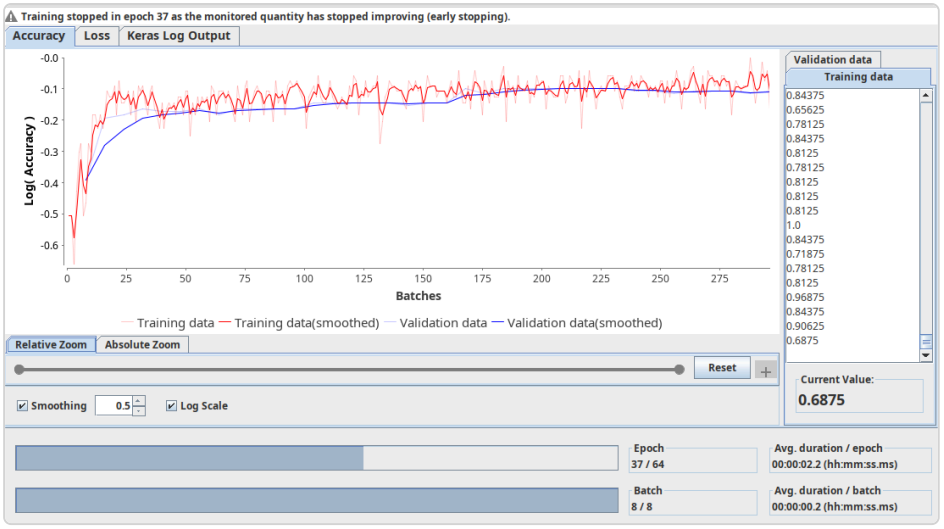

Figura 4: Acurácia - parada estratégica na época 37

Figura 5: Perda - parada estratégica na época 37

Resultados e Validação

O modelo atingiu uma acurácia aproximada entre 70% e 80% dentre várias combinações sobre os dados de validação, indicando boa capacidade de classificar corretamente pacientes diabéticos e não diabéticos. Adiante, seguimos com detalhes de uma das execuções finais da rede.

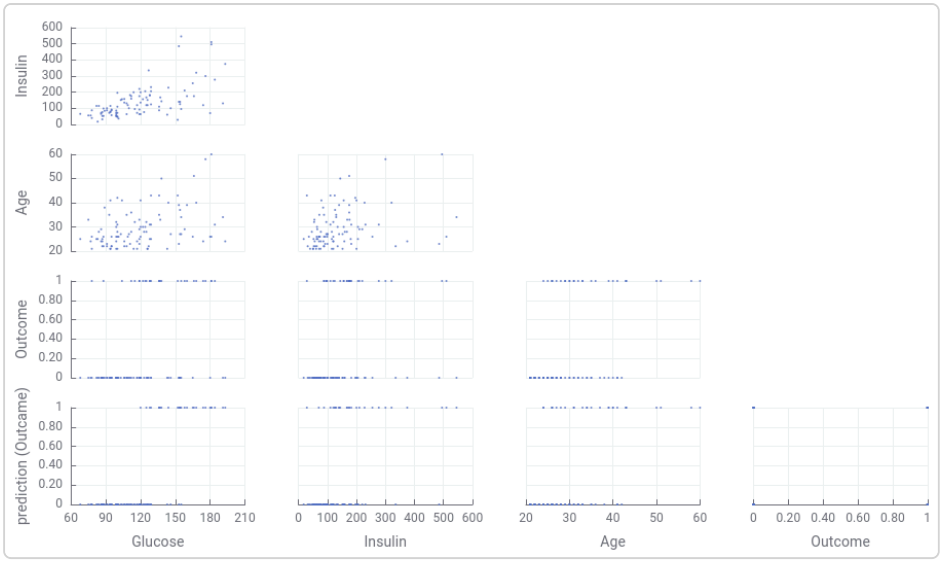

Figura 6: cruzamento das colunas mais relevantes

Análise dos Resultados do Modelo

O modelo apresentou os seguintes valores para as classes positivas e negativas, relacionados ao diagnóstico de diabetes:

Métrica	Classe 0 (negativo)	Classe 1 (positivo)	Valor Geral (Overall)
Verdadeiros Positivos (TP)	17	60	-
Falsos Positivos (FP)	10	11	-
Verdadeiros Negativos (TN)	60	17	-
Falsos Negativos (FN)	11	10	-
Recall (Sensibilidade)	0.607	0.857	-
Precision (Precisão)	0.630	0.845	-
Sensitivity (Sensibilidade)	0.607	0.857	-
Specificity (Especificidade)	0.857	0.607	-
F-measure (F1-Score)	0.618	0.851	-
Accuracy (Acurácia)	-	-	0.786
Cohen's Kappa[1]	-	-	0.469

Tabela 1: Scorer

Interpretação dos Resultados

1. **Acurácia (Accuracy) de 78,6%:**
   O modelo classificou corretamente 78,6% dos casos totais, o que é um resultado razoável para um problema complexo como a predição de diabetes. Este valor indica

uma boa capacidade do modelo em geral, mas é importante analisar outras métricas por causa do possível desequilíbrio entre as classes.

2. **Recall (Sensibilidade):**
   - Para a classe positiva (diabetes), o recall é alto (85,7%), indicando que o modelo conseguiu identificar a maioria dos pacientes que realmente têm diabetes, com poucos falsos negativos nessa classe. Isso é muito importante para não deixar casos de diabetes sem diagnóstico.
   - Já para a classe negativa, o recall é menor (60,7%), sinalizando que o modelo perdeu mais casos negativos (falsos negativos relativamente maiores).
3. **Precision (Precisão):**
   - Alta precisão para a classe positiva (84,5%) mostra que a maioria das previsões positivas realmente tinha diabetes, minimizando falsos positivos.
   - A precisão para a classe negativa é menor (63%), indicando uma confusão maior nas predições negativas.
4. **Specificity (Especificidade):**
   - Para a classe negativa, a especificidade é alta (85,7%), mostrando que o modelo é bom em reconhecer corretamente os negativos.
   - Para a classe positiva, a especificidade é menor (60,7%), confirmando que o modelo pode errar mais ao identificar falsos positivos nessa classe.
5. **F1-Score:**
   - O F1-score para a classe positiva (85,1%) indica equilíbrio entre precisão e recall, sendo forte na detecção correta da classe de interesse.
   - Para a classe negativa, F1 é menor (61,8%), reforçando que o modelo é mais eficiente para a classe com diabetes.
6. **Cohen’s Kappa:**
   Este índice mostra uma concordância moderada entre as previsões do modelo e os resultados reais (0,469), considerando a probabilidade de acerto ao acaso[1-1]. Apesar de não ser excelente, indica desempenho consistente acima do acaso.

Conclusão

Este trabalho apresentou a construção e avaliação de um modelo de rede neural para a predição da diabetes a partir de um conjunto de dados clínicos tabulares. A arquitetura adotada, combinada com técnicas de pré-processamento e ajuste de hiperparâmetros, permitiu alcançar resultados promissores.

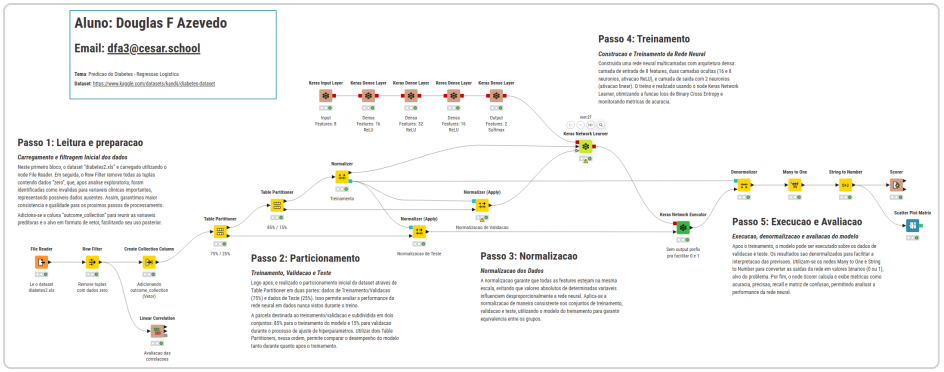

Figura 7: Workspace KNIME

Os resultados analisados por meio do nó Scorer evidenciaram que o modelo possui boa capacidade de identificar corretamente casos positivos de diabetes, com alta sensibilidade (recall de 85,7%) e precisão (84,5%) para essa classe. Essa característica é fundamental em contextos clínicos, pois minimiza a quantidade de diagnósticos falsamente negativos, reduzindo riscos à saúde dos pacientes.

Por outro lado, o desempenho para a classe negativa apresentou valores menores de recall (60,7%) e precisão (63%), indicando que o modelo ainda apresenta limitações na distinção de indivíduos não diabéticos. Isso ressalta uma oportunidade para melhorias, seja através de ajustes adicionais na arquitetura da rede, técnicas de balanceamento do conjunto de dados, ou otimização dos parâmetros de treino.

A acurácia global do modelo foi de 78,6%, com uma concordância moderada segundo o coeficiente de Cohen's Kappa (0,469), demonstrando que o modelo apresenta desempenho consistente acima do acaso, embora reconhecendo a necessidade de refinamentos para maior robustez.

Ademais, o uso do recurso de parada antecipada (early stopping) para controle do processo de treinamento, bem como o monitoramento criterioso das métricas de validação, são práticas essenciais para evitar o sobreajuste e garantir maior generalização do modelo.

Em suma, o modelo desenvolvido mostrou-se eficaz para o problema da predição da diabetes, especialmente na detecção correta da presença da doença, cumprindo seu propósito clínico principal. Para trabalhos futuros, recomenda-se explorar estratégias adicionais de otimização, inclusão de novas fontes de dados, e análise aprofundada para aprimorar a sensibilidade e precisão equilibradas entre as classes.

---

1. [1] Cohen’s Kappa avalia a concordância entre rótulos preditos e verdadeiros com correção para acertos ao acaso, variando de -1 (discordância total) a 1 (concordância perfeita). Valores entre 0.4 e 0.6 indicam concordância moderada. ↩ ↩