Universidade de Évora

Escola de Ciências e Tecnologia

Previsão de Incumprimento de Crédito
Uma Abordagem Baseada no Algoritmo K-Nearest Neighbors (KNN)

Trabalho realizado por:
Madalena Marques

Unidade Curricular: Tópicos de Inteligência Artificial e Ciências de Dados
Docente: Professor Doutor Luis Rato
Ano Letivo: 2025/2026

Évora, 23 de janeiro de 2026

Contents

0.1 Introdução

No âmbito da unidade curricular de Tópicos de Inteligência Artificial e Ciências Dados, foi proposto a realização de um trabalho final com objetivo de aplicar os conteúdos lecionados ao longo das aulas.

O dataset utilizado para este estudo é *"Credit Card Default (UCI)"* procurado no repositório *UCI Machine Learning* que retrata conjunto de dados proveniente de uma instituição financeira de Taiwan, que contém informações detalhadas de 30 000 clientes, incluindo variáveis demográficas (como género, escolaridade, idade e estado civil), o histórico de limites de crédito concedidos (LIMIT BAL), o registo de pagamentos e os montantes de faturação mensal (BILL AMT) num período de seis meses. Para este estudo em questão foi usado programa software estatístico *RStudio* cujo os output e os resultados serão apresentados e analisados ao longo do trabalho sobre a forma de gráficos e tabelas. Além disso foram usados diversos pacotes do R para facilitar a manipulação de dados e a criação dos gráficos.

O relatório encontra-se dividido em onze partes em que inicialmente são referidas abordagens deste tema e questões éticas e legais dos dados. Procede-se a uma análise descritiva das variáveis e da relação entre elas. De seguida é analisada a qualidade dos dados e realizadas ações para resolver possíveis problemas para a construção modelo preditivo. Após a limpeza dos dados procede-se à construção do modelo preditivo KNN e avaliação da sua qualidade de predição. E por último, é salvo o modelo para *deployment* e discutido os procedimentos de manutenção do modelo em médio prazo.

0.2 Problema e o contexto do *dataset*

A concessão de crédito é uma ferramenta fundamental para funcionamento da economia moderna, permitindo o consumo e o investimento tanto por parte de pessoas como de empresas.

No entanto, a gestão de risco é o maior desafio das instituições financeiras, uma vez que o incumprimento no pagamento das faturas de crédito pode levar a perdas financeiras significativas. Tal como a recomendação de cupões exige identificar o perfil certo para evitar o desinteresse, na área financeira, identificar antecipadamente quais os clientes que têm maior probabilidade de entrar em incumprimento é crucial. Perfis diferentes (com base na idade, escolaridade e histórico de pagamentos) apresentam comportamento de risco distintos.

A melhor a abordagem para tratar este problema seria *machine learning* surge como uma ferramenta essencial para criar sistemas de previsão de risco mais precisos e eficientes. Através da análise de dados históricos, é possível treinar modelos para detetar padrões que humanos poderiam ignorar. O objetivo central é prever o incumprimento no mês seguinte. Ao prever com sucesso se um cliente irá falhar o pagamento, a instituição pode tomar medidas preventivas, como ajustar limite de crédito ou oferecer planos de pagamento, aumentando a sua rentabilidade e evitando custos de recuperação de dívida.

0.3 Questões Éticas, legais e de Fairness

A utilização deste *dataset* de clientes bancários exige conformidade estrita com o Regulamento Geral sobre a Proteção de Dados (RGPD). Este *dataset* utiliza um rótulo, essencial para garantir que nenhum cruzamento de variáveis que permita a identificar a pessoa, devem ser utilizadas apenas as variáveis estritamente necessárias para a previsão do risco. Além disso, no setor financeiro, o cliente tem direito de saber por que razão o seu crédito foi negado ou por que foi classificado como perfil de risco. Isto exige que o modelo seja interpretável. Para mitigar os riscos anteriores proponho que se realizem teste de performance do modelo separados por género e educação para verificar se o erro é significativamente maior num grupo do que noutro, também documentar como cada variável influência a decisão final do modelo KNN e por último é necessária manutenção regular para garantir que o seu desempenho se mantenha ao longo do tempo.

0.4 Análise Descritiva dos dados

Foi feita uma análise descritiva sob forma de gráficos de barras, histogramas e caixas de bigode para cada variável do dataset.

0.4.1 Sex

A variável sex é categórica, definida como proporção de incumprimento do crédito por género do cliente em que 1 representa o sexo masculino e 2 o sexo feminino.

A figura 1 mostra análise descritiva desta variável sob forma de gráfico de barras.

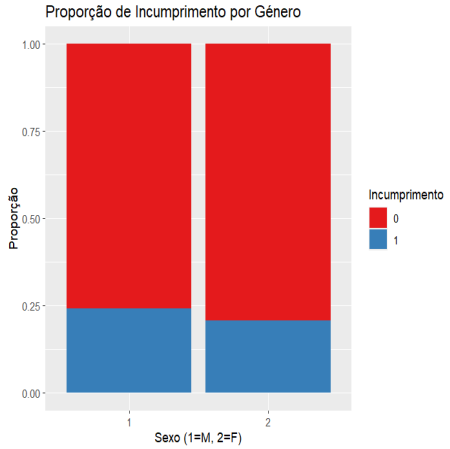

Figure 1: Análise descritiva da variável sex

Verificou-se uma diferença visível na taxa de incumprimento. A barra azul no grupo 1 (Homens) é ligeiramente superior à do grupo 2 (Mulheres).

0.4.2 Education

A variável education é categórica, definida como distribuição da população por escolaridade em que 1 representa Pós-graduação, o 2 a Universidade, 3 o Ensino Médio e 4 os outros.

A figura 2 mostra análise descritiva desta variável sob forma de gráfico de barras.

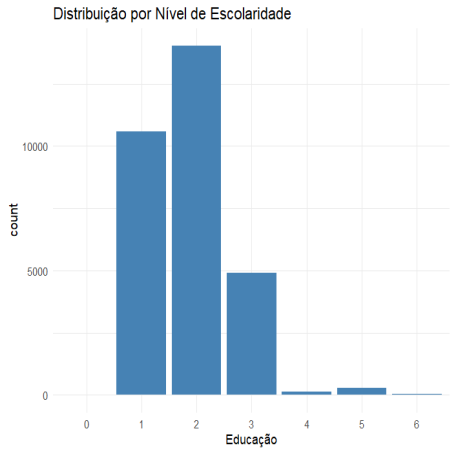

Figure 2: Análise descrita da variável education

Verificou-se um maior aumento na população com escolaridade a nível de ensino superior do que em população com nível de escolaridade mais baixo a nível de crédito bancário.

0.4.3 Marriage

A variável marriage é categórica, definida como proporção de incumprimento por estado civil em que 1 representa os Casados, 2 os Solteiros, 3 os Outros.

A figura 3 mostra análise descritiva desta variável sob forma de gráfico de barras.

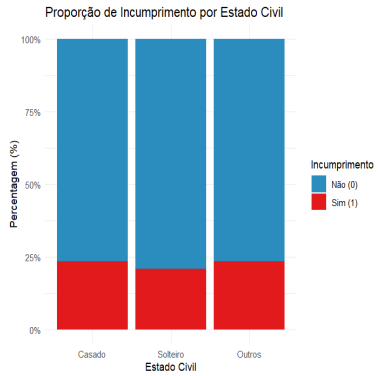

Figure 3: Análise descrita da variável marriage

Verificou-se que os clientes solteiros apresentam uma taxa de incumprimento inferior (21%) face aos casados (24%).

0.4.4 LIMIT Bal

A variável LIMIT Bal é numérica, definida como montante do crédito concedido.

A figura 4 mostra análise descritiva desta variável sob forma de histograma.

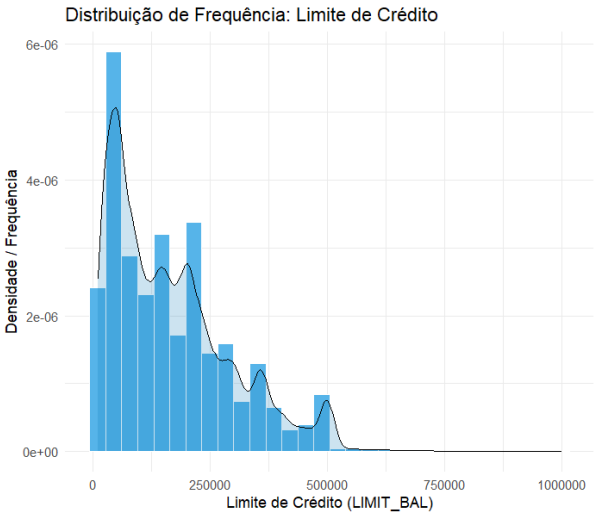

Figure 4: Análise descrita da variável LIMIT Bal

Verificou-se que a concessão de crédito é fortemente assimétrica, com uma concentração predominante em valores de crédito mais baixo.

0.4.5 AGE

A variável Age é numérica, definida como idade do cliente (em anos).

A figura 5 mostra análise descritiva desta variável sob forma de histograma.

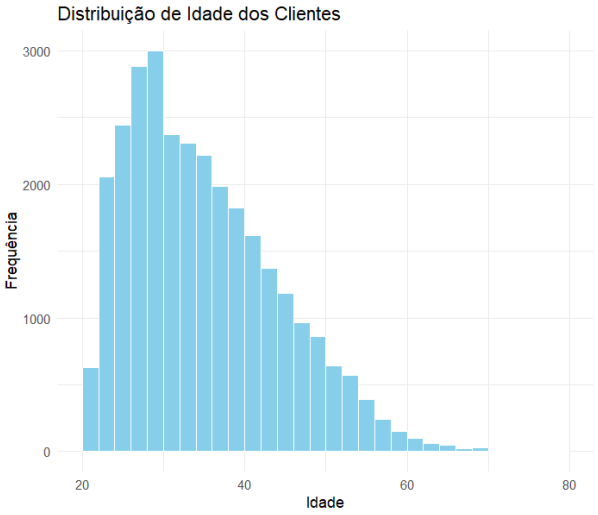

Figure 5: Análise descrita da variável age

Verificou-se uma distribuição unimodal com uma ligeira assimetria positiva. A maior densidade de clientes da instituição financeira situa-se na faixa dos jovens adultos, especificamente entre os 25 e os 35 anos.

0.4.6 BILL-AMT1-6

A variável BILL-AMT1-6 é numérica, definida como valor da fatura mensal entre abril a setembro .

A figura 6 mostra análise descritiva desta variável sob forma de caixa de bigodes.

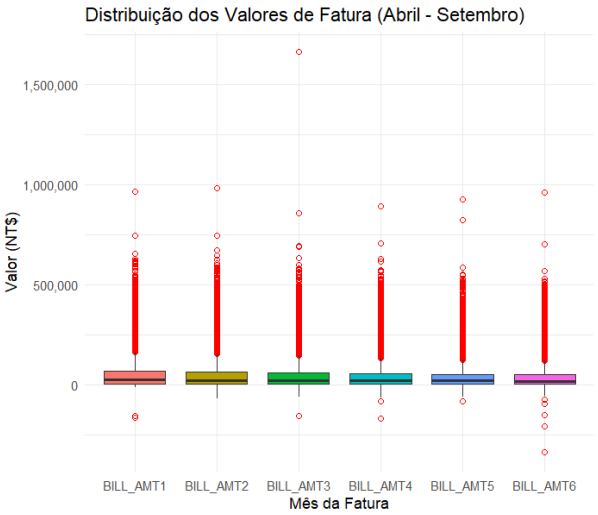

Figure 6: Análise descrita da variável BILL-AMT1-6

Verificou-se uma elevada concentração de clientes com faturas reduzidas, contrastando com uma minoria de clientes de 'alto valor' cujos gastos ultrapassam consistentemente os 500 000.

0.4.7 Default payment next month(Y)

A variável default payment next month(Y) é categórica, indicando se o cliente entrou em incumprimento em que 1 representa que sim e 0 que não .

A figura 7 mostra análise descritiva desta variável sob forma de gráfico de barras.

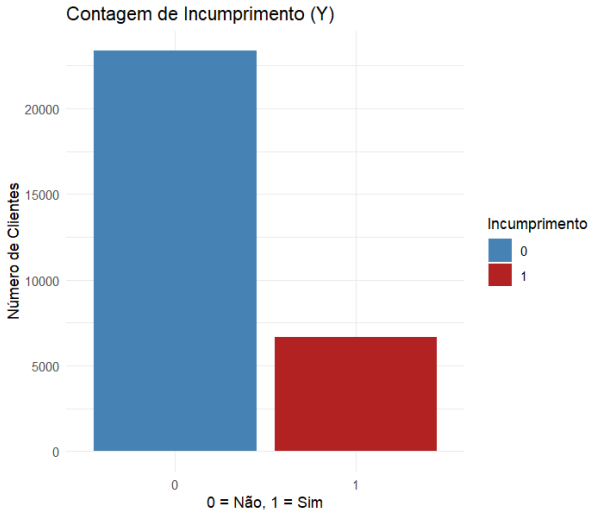

Figure 7: Análise descritiva da variável default payment next month(Y)

Verificou-se um acentuado desequilíbrio entre classes, com a vasta maioria da amostra a pertencer ao grupo de cumpridores (0). Indicando que a população cumpre os prazos de pagamento do crédito.

0.5 Relação entre variáveis

Aqui pretendemos encontrar relações importantes entre variáveis duas-a-duas, neste caso, da estado do pagamento recente, incumprimento, limite de crédito e a escolaridade.

0.5.1 Estado do pagamento recente vs Incumprimento

A análise do figura 8 revela uma correlação positiva extremamente forte entre o atraso nos pagamentos e a probabilidade de incumprimento. É evidente que o comportamento em setembro serve como um sinal de alerta precoce: a partir dos dois meses de atraso, o perfil de cliente altera-se drasticamente, tornando o incumprimento o cenário mais provável (superior a 70%).

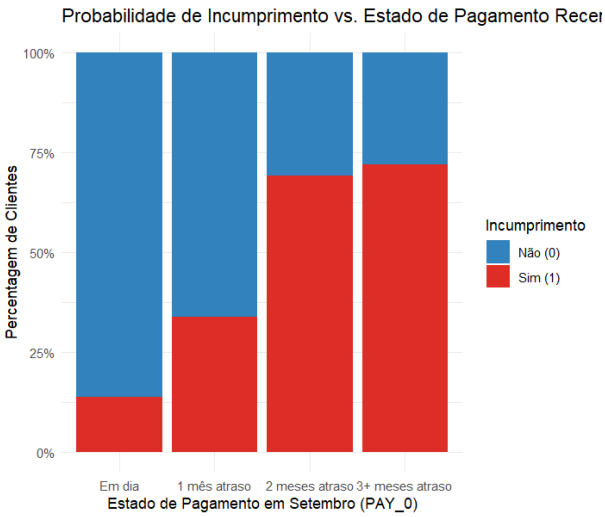

Figure 8: Relação do estado do pagamento recente e o incumprimento sob forma de gráfico barras

0.5.2 Limite de Crédito vs Incumprimento

A análise do figura 9 revela uma tendência clara: clientes com limites de crédito mais elevados apresentam uma menor propensão para incumprimento. A mediana do grupo cumpridor situa-se significativamente acima do grupo em *default*, o que confirma que o *planfond* atribuído é um forte indicador de solvabilidade.

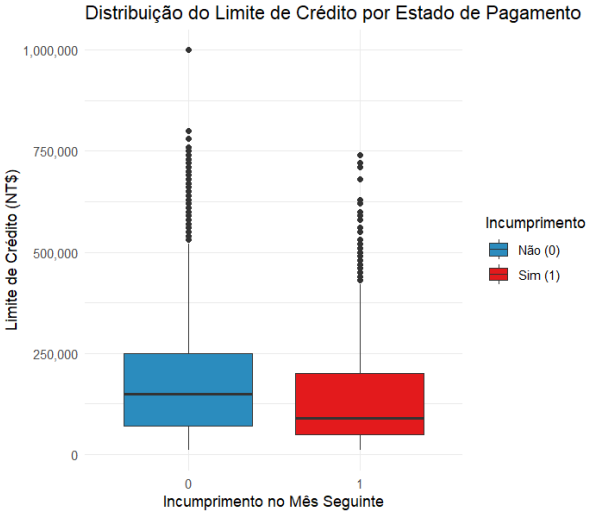

Figure 9: Distribuição do limite de crédito por estado pagamento sobre a forma de caixa de bigodes

0.5.3 Correlação entre faturas

A análise da correlação entre as variáveis de faturação revelou a existência de uma forte multicolinearidade, com os coeficientes de person frequentemente superiores a 0.90 entre meses consecutivos. Esta observação sugere que o endividamento dos clientes é persistente ao longo do semestre analisado.

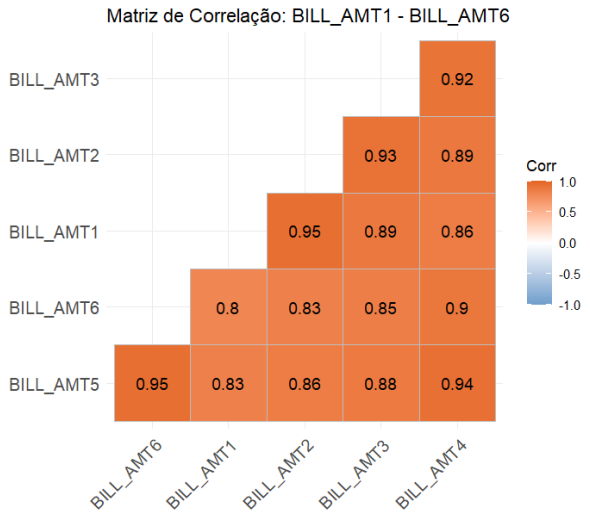

Figure 10: Matriz correlação entre faturas

0.5.4 Escolaridade vs Incumprimento

A figura 11 utiliza barras empilhadas a 100%, para comparar a proporção de pessoas cumpridoras (azul) e incumpridoras (vermelho) em diferentes níveis de instrução. Observa-se que o risco de incumprimento tende a aumentar à medida que o nível de escolaridade formal diminui entre as categorias principais. A pós-graduação, apresenta a menor a taxa de incumprimento entre os níveis académicos, situando-se em cerca de 20%. Já os universitários, a taxa sobe ligeiramente para aproximadamente 24%. No ensino médio, regista a taxa mais elevada de incumprimento entre estas três categorias, atingindo cerca de 25-26%. E por último, na categoria "outros" apresenta a taxa de risco mais baixa (inferior a 10%), o que pode indicar um grupo muito específico ou a necessidade de uma limpeza de dados mais profunda.

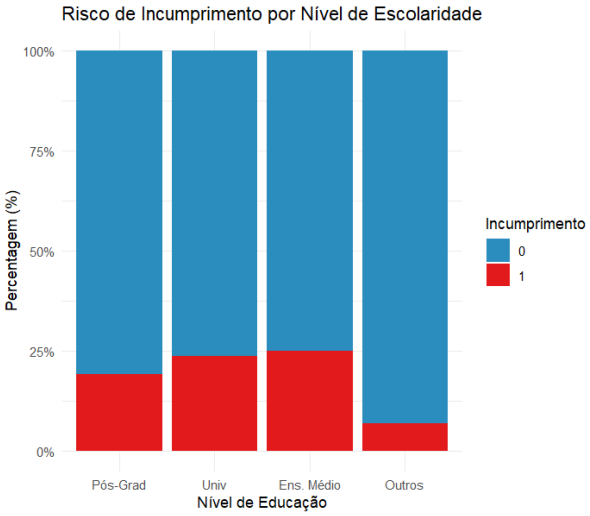

Figure 11: -Relação entre a escolaridade e o incumprimento sob forma de gráfico de barras

0.6 Qualidade dos dados

Para esta análise, foram usadas algumas funções para procurar valores que correspondessem.

0.6.1 Duplicados, valores em falta, valores inconsistentes ou inválidos

Duplicados

Aqui queremos encontrar o número de linhas que tenham valores iguais para todas as variáveis. Para este fim, foi utilizado o método *duplicated()* para identificar valores duplicados e o método para contar quantos existem. A figura 12 mostra o número de duplicados deste *dataset*.

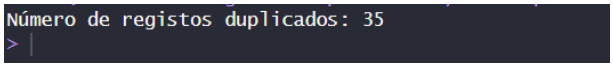

Figure 12: Número de duplicados

Valores em falta (NAs)

Procuram-se valores que estejam vazios ou nulos. Neste caso, foi aplicada uma função de diagnóstico em todas as colunas do *dataset* para quantificar a ausência de dados. Não foram encontrados *missing values* em nenhumas das colunas. A figura 13 mostra output gerado para cálculo dos *missing values*.

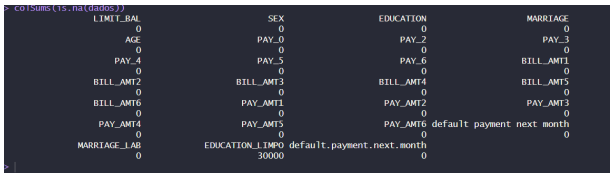

Figure 13: Missing Values

Valores inconsistentes ou inválidos

Valores que não façam sentido com os restantes. A partir da análise descritiva foi possível encontrar valores inconsistentes em 2 variáveis: *education* e *marriage*.

0.7 Divisão Train/Test e Replicabilidade

Para evitar o *overfitting*, o *dataset* de 30 000 clientes não deve ser usado na totalidade para treinar o modelo. Deve ser dividido em duas partes:

1. *Training set*: que vai ser usado pelo KNN para aprender os padrões de vizinhança e as características dos clientes.

2. *Test set*: Dados "invisíveis" do modelo, usados apenas para avaliar a sua capacidade real de prever o incumprimento em novos clientes.

Para tal foi segmentado em dois conjuntos independentes: Treino (75%) e Teste (25%). Esta divisão é crucial para validar a capacidade de generalização do algoritmo KNN perante novos dados de clientes. Visando assegurar a replicabilidade total da experiência foi fixada uma variável aleatória, permitindo que os resultados reportados possam ser reproduzidos fielmente.

Adicionalmente, aplicou-se uma partição estratificada para garantir que a distribuição da variável alvo (incumprimento) se mantém consistente em ambos os subconjuntos, prevenindo alterações na performance preditiva do modelo. O desenvolvimento deste estudo foi realizado no programa software estatisticos *R (RStudio)*. Todos os pacotes necessários à sua execução encontram-se explicitamente declarados no início do script, assegurando a reprodutibilidade dos resultados e a independência face a configurações específicas de ambientes externos.

0.8 Estratégia de Limpeza e Preparação

Limpeza Estrutural

Verificar e eliminar os duplicados para evitar o *overfitting* do modelo. Confirmar a totalidade dos 30 000 registos. Caso existissem valores nulos, seriam aplicadas técnicas de imputação, uma vez que KNN não suporta entradas vazias.

Limpeza Semântica (Consistência de Domínio)

Corrigir os valores não documentados identificados na análise exploratória.

Na variável educação, agrupar os valores 0, 5 e 6 na categoria 4 ("outros"). E na variável estado civil, mapear o valor 0 para categoria 3 ("Outros"). Garantir que o KNN agrupe vizinhos com base em categorias com significado real, evitando a criação de *"clusters"* de erro.

0.9 Modelo KNN

O modelo preditivo foi implementado recorrendo ao algoritmo KNN, cuja lógica de classificação assenta na identificação de perfis de risco semelhantes através do cálculo de distâncias euclidianas entre observações, a escala das variáveis preditoras assume um papel crítico. Neste conjunto de dados, variáveis financeiras como LIMIT BAL e BILL AMT apresentam valores que atingem a ordem do milhão, enquanto variáveis demográficas e de estado civil (SEX, MARRIAGE) variam num intervalo restrito entre 1 e 6. Sem um tratamento prévio, a magnitude das variáveis financeiras dominaria o cálculo da distância, tornando o modelo 'cego' aos fatores demográficos. Para garantir que todas as variáveis contribuam de forma igual para a definição de vizinhança, foi aplicado um processo de *Standardization (Z-score normalization)* através das operações de *center* (subtração da média) e *scale* (divisão pelo desvio padrão).

A seleção do k para este estudo foi realizada recorrendo a *k-fold* sobre o conjunto de treino, garantindo uma estimativa robusta da performance do modelo e evitando *overfitting*, utilizando a validação cruzada com 10 *folds* que assegura a distribuição da variável alvo (incumprimento) se manteve consistente em cada partição. Para cada valor de k testado, o modelo foi treinado em 9 *folds* e validado no *fold* restante, repetindo o processo até que todos os subconjuntos tivessem sido utilizados como validação, sendo que o critério de seleção é maximização da métrica AUC-ROC, considerada mais adequada em contextos de classes desequilibradas.

Importa salientar que todo o processo de validação cruzada, incluindo o pré-processamento (*center e scale*), foi realizado exclusivamente no conjunto de treino, sendo o conjunto de teste reservado apenas para a avaliação final do modelo, prevenindo qualquer forma de data *leakage*.

Conforme ilustrado no figura 14 que representa a otimização do modelo, o valor de k=25 foi selecionado por maximizar a exatidão do modelo (79.03%). Este valor garante que o modelo ignore ruídos estatísticos e capture as tendências comportamentais reais, como a disparidade de risco observada entre diferentes níveis de escolaridade. A análise da matriz de confusão, representada na figura 15 revela um modelo excessivamente otimista e arriscado, com elevado índice de falsos negativos (1717 casos). O baixo valor de Kappa (0.1378) confirma que o algoritmo está a ser dominado pela classe maioritária dos que cumpre o pagamento, falhando na identificação de perfis de risco, mesmo com a inclusão de variáveis importantes como o nível de escolaridade.

Para uma aplicação real, seriam necessárias técnicas adicionais de balanceamento de dados ou ajuste de limiares de decisão para priorizar a sensibilidade face ao incumprimento.

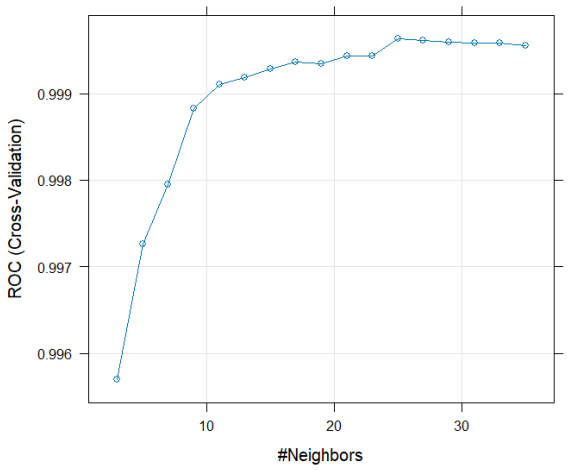

Figure 14: Gráfico de Otimização do modelo KNN

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6873 1717
         1  170  240

               Accuracy : 0.7903
                 95% CI : (0.7818, 0.7987)
    No Information Rate : 0.7826
    P-Value [Acc > NIR] : 0.03744

                  Kappa : 0.1378

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9759
            Specificity : 0.1226
         Pos Pred Value : 0.8001
         Neg Pred Value : 0.5854
             Prevalence : 0.7826
         Detection Rate : 0.7637
   Detection Prevalence : 0.9544
      Balanced Accuracy : 0.5492

       'Positive' Class : 0
```

Figure 15: Confusion Matrix do modelo KNN

0.10 Avaliação e *Fairness*

O modelo KNN implementado com k=25 atingiu uma accuracy de 77.77% no conjunto de teste, esta métrica isolada é insuficiente para validar a eficácia do modelo. A análise mais profunda revelou que o valor de Kappa (0.087) e a especificidade de 8.53% indicam que o modelo falha criticamente na identificação de clientes em incumprimento.

Além disso, o AUC de 0.227 na curva PR, como ilustrado na figura 16, confirma que o modelo tem grande dificuldade em equilibrar a precisão e a recuperação da classe minoritária. A avaliação imparcial demonstra que o modelo, na sua configuração atual, é "cego" às variações de risco entre subgrupos. Embora existam diferenças reais na taxa de incumprimento por nível de escolaridade — com o "Ensino Médio" a apresentar maior risco que a "Pós-Graduação" — o modelo tende a classificar quase todos os perfis como cumpridores para maximizar a exatidão.

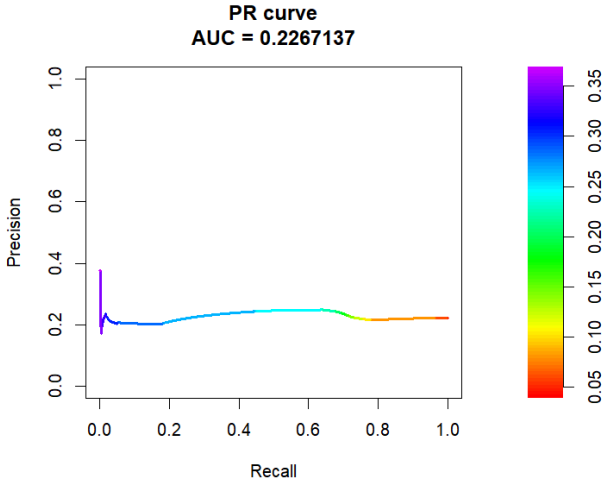

Figure 16: Gráfico de Curva PR para modelo KNN com k=25

Após o treino e validação do modelo KNN, procedeu-se ao seu armazenamento de forma a garantir reutilização, reprodutibilidade e preparação para *deployment*. O modelo final KNN (k = 25) foi guardado em conjunto com todo o pipeline de pré-processamento que inclui a normalização das variáveis numéricas *(Z-score)*, *encoding* das variáveis categóricas, assegurando que novos dados são tratados de forma consistente com os dados de treino.

O objeto do modelo foi sequenciado em formato RDS, permitindo o seu carregamento posterior sem necessidade de re-treino.

0.11 Manutenção

A manutenção do modelo KNN (k=25) é vital para mitigar a degradação da sua performance preditiva (*Concept Drift*). Dado que o modelo apresenta atualmente uma reduzida capacidade de deteção de incumprimento (Especificidade de 8,53% e AUC-PR de 0,227), a monitorização deve focar-se na taxa de Falsos Negativos para evitar prejuízos financeiros acumulados. Além da performance técnica, a manutenção incluirá auditorias imparciais para assegurar que as decisões de crédito não penalizam desproporcionalmente subgrupos vulneráveis. Recomenda-se um ciclo de re-treino semestral, incorporando novos dados e técnicas de balanceamento de classes para melhorar a robustez e a justiça do sistema de apoio à decisão.

0.12 Conclusão

O presente trabalho permitiu o desenvolvimento e a avaliação de um sistema de apoio à decisão baseado no algoritmo *K-Nearest Neighbors (KNN)* para a previsão de incumprimento de crédito. Através de uma

metodologia rigorosa que abrange desde a análise exploratória até à validação do modelo, foi possível extrair conclusões críticas sobre a aplicação de Inteligência Artificial no setor financeiro.

A otimização do parâmetro para k=25 permitiu encontrar um equilíbrio na vizinhança estatística, contudo, os resultados obtidos sublinham a complexidade do problema. Embora a *acurracy* global possa parecer elevada, métricas mais robustas como o *Kappa* de 0.087 e a AUC-PR de 0.227 revelam que o modelo enfrenta dificuldades significativas devido ao desequilíbrio de classes.

A reduzida especificidade de 8.53% indica que o algoritmo é excessivamente conservador, falhando na identificação da maioria dos clientes em risco de incumprimento, o que teria impactos financeiros diretos numa instituição bancária real.

Além disso, a disparidade de escalas entre variáveis financeiras (como o LIMIT BAL) e demográficas exigiu uma normalização criteriosa para evitar previsões tendenciosas. No âmbito da ética e do RGPD, a análise de *fairness* evidenciou que, apesar de o modelo ser tecnicamente "cego" a subgrupos, as disparidades históricas na concessão de crédito podem perpetuar discriminações se não forem ativamente monitorizadas. A igualdade algorítmica deve, portanto, ser um pilar tão importante quanto a eficácia preditiva.

Em suma, o projeto cumpriu os objetivos pedagógicos ao demonstrar que a implementação de modelos de Inteligência artificial exige uma visão global, onde a preparação dos dados, a escolha das métricas de avaliação e a vigilância ética são fundamentais para garantir um sistema de crédito justo e financeiramente sustentável.

0.13 Referências bibliográficas

GDPR2016, author = European Parliament and Council, title = Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), year = 2016, howpublished = Official Journal of the European Union, note = OJ L 119, 4.5.2016, p. 1–88, url = https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

Fox, J (2002). An R and S-Plus Companion to Applied Regression. Sage Publications.