Life Insurance

José Filipe[1] and José Cunha[2]

[1] Universidade de Coimbra
josefilipe1@gmail.com
[2] Universidade de Coimbra
josecunha.8989@gmail.com

Abstract. Este projeto explora a aplicação de modelos probabilísticos para apoiar a tomada de decisão no contexto da aceitação de seguros de vida. Implementámos e comparamos dois métodos: um modelo Bayesiano e uma regra determinística de pontuação, ambos utilizados para classificar indivíduos como aceites ou rejeitados com base em variáveis pessoais, clínicas e económicas. O modelo Bayesiano, ao estimar probabilidades posteriores, permite uma abordagem mais flexível e transparente, incorporando a incerteza no processo de decisão. Ambos os métodos foram avaliados num conjunto de dados real, com especial foco no desempenho, interpretabilidade e robustez. Desenvolvemos ainda uma aplicação interativa em Streamlit que permite simular a decisão para diferentes perfis de clientes. Os resultados obtidos indicam que o modelo Bayesiano supera a regra determinística em termos de precisão e fiabilidade, revelando-se uma alternativa promissora para sistemas de decisão baseados em risco no setor segurador.

Keywords: Modelos Probabilísticos · Decisão em Seguros de Vida · Classificação Bayesiana

Introdução

A tomada de decisão sobre a aquisição de um seguro de vida depende de diversos fatores de natureza pessoal, clínica e económica. Esta decisão envolve frequentemente incerteza, tornando os métodos baseados em probabilidade uma escolha natural para apoiar indivíduos nesse processo. Neste projeto, explora-se a aplicação de técnicas de fusão de informação para construir um sistema de apoio à decisão que estime, com base em dados históricos, a probabilidade de um indivíduo optar por adquirir um seguro de vida.

A abordagem seguida baseia-se num modelo Bayesiano, treinado com exemplos de decisões previamente tomadas por indivíduos. Através da inferência probabilística, o modelo permite estimar, para novos casos, a probabilidade de recomendação da compra do seguro, com base em variáveis como idade, estado civil, número de dependentes, estado de saúde e salário mensal.

O trabalho inclui também o desenvolvimento de uma aplicação interativa que facilita a análise exploratória dos dados, a visualização de métricas e a geração de previsões personalizadas. Este relatório descreve o processo de modelação, os resultados obtidos e a avaliação do desempenho do sistema proposto.

Descrição dos dados

O conjunto de dados utilizado neste projeto consiste em registos individuais que incluem variáveis associadas ao perfil pessoal, clínico e económico de cada pessoa, juntamente com a respetiva decisão de adquirir ou não um seguro de vida. Estas decisões são consideradas corretas e servem como base para o treino supervisionado do modelo.

Variável	Descrição	Tipo	Valores possíveis
`Gender`	Género	Categórica	0 = Feminino, 1 = Masculino
`Age`	Idade	Contínua	Inteiros entre 34 e 101
`MaritalStatus`	Estado civil	Categórica	0 = Solteiro, 1 = Casado
`Dependents`	Nº de dependentes	Ordinal	0, 1, 2, 3 (≥ 3)
`PhysicalStatus`	Atividade física	Ordinal	0 = Sedentário, 1 = Moderado, 2 = Ativo
`ChronicDiseases`	Doenças crónicas	Ordinal	0 = Nenhuma, 1 = Moderada, 2 = Severa
`MonthlySalary`	Salário mensal	Contínua	Entre 1370€ e 3800€
`Decision`	Decisão de seguro	Binária	0 = Não, 1 = Sim

Table 1. Descrição das variáveis do conjunto de dados

Adicionalmente, foi calculada uma pontuação de risco para cada indivíduo com base numa regra determinística fornecida por uma seguradora, que agrega contributos das variáveis Age, ChronicDiseases,

MonthlySalary e Dependents. Esta pontuação é utilizada posteriormente como referência comparativa ao desempenho do modelo Bayesiano, mas não é usada como variável de entrada durante o treino.

A distribuição das variáveis demonstra boa diversidade nos perfis representados, permitindo explorar relações entre fatores de risco e decisões de contratação. Os dados foram utilizados sem alterações nos seus significados originais, sendo apenas aplicados filtros interativos na interface para facilitar a análise. É possível aceder à interface interativa a partir do seguinte url `https://josepedrocunhazzz-bayesian-lifeinsurance-da` `streamlit.app`.

Modelo Proposto

O modelo desenvolvido para este projeto baseia-se na formulação clássica do classificador Naive Bayes, que assume independência condicional entre as variáveis explicativas dado o resultado. A decisão de um indivíduo (Decision) é considerada como a variável de interesse, e todas as outras variáveis são tratadas como evidência observável. Além da distribuição normal, foi também considerada uma modelação alternativa para variáveis contínuas com recurso a Misturas Gaussianas

Modelo Bayesiano Proposto

O modelo desenvolvido para este projeto baseia-se na formulação clássica do *classificador Naive Bayes*, que assume independência condicional entre as variáveis explicativas dado o resultado. A decisão de um indivíduo (`Decision`) é considerada como a variável de interesse, e todas as outras variáveis são tratadas como evidência observável.

O objetivo do modelo é estimar a probabilidade posterior $P(T \mid X_1, X_2, \ldots, X_n)$, onde $T \in \{0, 1\}$ representa a decisão de contratar (1) ou não contratar (0) o seguro de vida, e X_i são as variáveis observadas.

Estrutura do Modelo

Para cada observação, a probabilidade de cada classe é estimada através da regra de Bayes:

$$P(T \mid X) \propto P(T) \prod_{i=1}^{n} P(X_i \mid T)$$

onde:

- $P(T)$ é a probabilidade *a priori* da decisão (calculada a partir dos dados),
- $P(X_i \mid T)$ são as probabilidades condicionais de cada variável dado o valor da decisão.

Tratamento das Variáveis

- **Variáveis categóricas e discretas** (ex.: `Gender`, `MaritalStatus`, `Dependents`) foram tratadas com tabelas de frequências relativas por classe, aplicando *suavização de Laplace* para evitar probabilidades nulas.
- **Variáveis contínuas** (`Age`, `MonthlySalary`) foram modeladas assumindo uma *distribuição normal* para cada classe, estimando a média e o desvio padrão com base nos dados disponíveis:

$$P(X_i = x \mid T = t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(x - \mu_t)^2}{2\sigma_t^2}\right)$$

Inferência e Decisão

A decisão final para cada indivíduo é tomada pela *regra do máximo a posteriori* (MAP):

$$\hat{T} = \arg\max_{t \in \{0,1\}} P(T = t \mid X)$$

Esta abordagem permite estimar não só a classe mais provável, mas também a probabilidade associada a cada decisão, o que é particularmente útil em contextos de risco.

Implementação e Interface Interativa

Para tornar o modelo acessível e interativo, foi desenvolvida uma aplicação utilizando a biblioteca `Streamlit`, permitindo a exploração dos dados, a visualização de métricas e a geração de previsões personalizadas em tempo real. A interface é composta por cinco separadores principais, cada um com funcionalidades específicas.

Carregamento e Pré-processamento dos Dados

Os dados são carregados a partir do ficheiro `lifeInsurance.txt`, sendo automaticamente processados e filtrados conforme os critérios definidos pelo utilizador na barra lateral. É também calculada, para cada indivíduo, a *pontuação de risco* com base na regra da seguradora, permitindo futuras comparações.

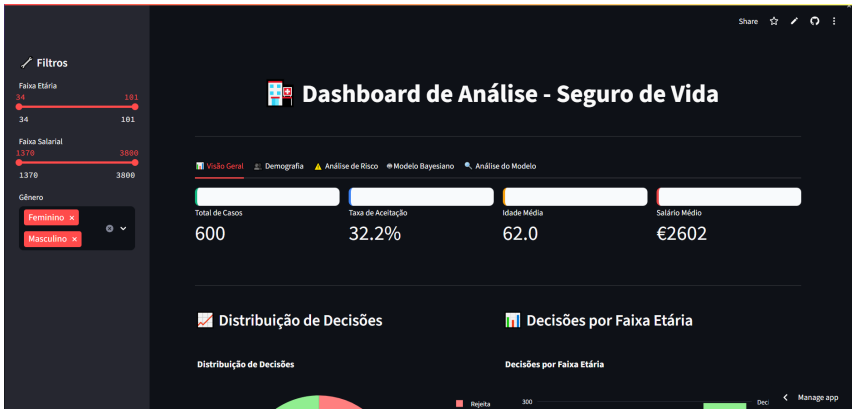

Fig. 1. Exemplo ilustrativo da interface interativa

Separadores da Aplicação

– **Visão Geral**: Apresenta métricas agregadas (ex: total de casos, idade média, salário médio), e gráficos com a distribuição de decisões e decisões por faixa etária.

– **Demografia**: Oferece visualizações relacionando variáveis como idade, salário e número de dependentes com as decisões tomadas. Inclui ainda uma matriz de correlação entre variáveis relevantes.

– **Análise de Risco**: Foca-se exclusivamente na análise da **regra dos 50 pontos** da seguradora, mostrando a distribuição dos scores de risco e sua relação com a decisão real.

– **Modelo Bayesiano**: Permite ao utilizador:

- Treinar o modelo Naive Bayes com os dados filtrados;
- Visualizar métricas de desempenho como accuracy e matriz de confusão;
- Utilizar um formulário para *inserir manualmente os atributos de um indivíduo* e obter a predição do modelo, incluindo probabilidade e score de risco.

– **Análise do Modelo**: Reúne uma análise aprofundada sobre:

- Performance do classificador;
- Comparação com a regra dos 50 pontos;
- Seleção de variáveis;
- Impacto de discretizar variáveis contínuas;
- Testes de normalidade;
- Recomendações para melhorias futuras.

Experiência do Utilizador

A aplicação foi desenhada com foco na clareza visual e facilidade de navegação. Elementos visuais como **gráficos interativos com Plotly**, **indicadores com código de cor** e **textos explicativos dinâmicos** tornam a experiência informativa e intuitiva, permitindo que tanto perfis técnicos como não técnicos possam explorar os resultados.

Avaliação e Resultados

A avaliação do modelo Bayesiano foi conduzida através de métricas clássicas de classificação, com base na comparação entre as previsões geradas pelo modelo e as decisões reais observadas nos dados. Adicionalmente, foi realizada uma análise comparativa com a regra determinística da seguradora, que serve como linha de base (*baseline*).

Desempenho do Modelo

A accuracy global do modelo Naive Bayes treinado com todas as variáveis disponíveis foi de aproximadamente **84%**. A matriz de confusão indica uma boa capacidade de distinguir entre os casos positivos (compra) e negativos (não compra), com um número elevado de verdadeiros positivos e verdadeiros negativos.

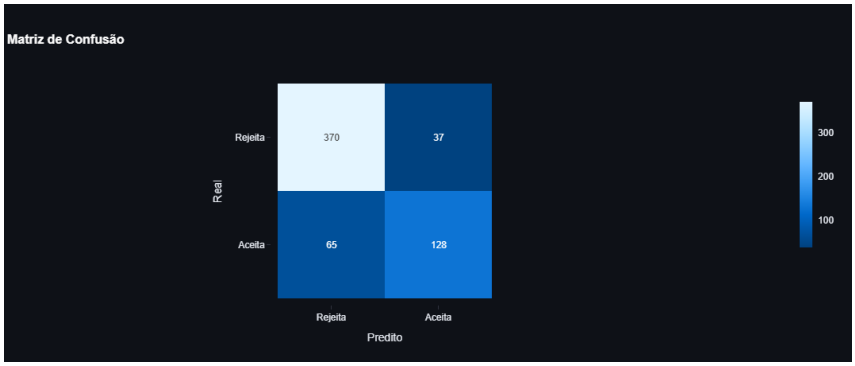

Fig. 2. Matriz de confusão do modelo

Além da accuracy, foram analisadas outras métricas relevantes:

- **Recall** para a classe positiva ("Aceita") superior a 90%, indicando forte capacidade do modelo em identificar corretamente os indivíduos que devem contratar o seguro.
- **Precisão** para a classe negativa ("Rejeita") elevada, revelando confiança nos casos em que o modelo não recomenda a aquisição do seguro.

Comparação com a Regra dos 50 Pontos

A regra da seguradora atribui pontos a quatro variáveis (idade, doenças crónicas, salário e dependentes) e recomenda a compra do seguro caso a pontuação ultrapasse os 50 pontos.

Modelo	accuracy obtida
Naive Bayes (7 variáveis)	˜84%
Regra dos 50 pontos	˜75%

A comparação mostra que o modelo Bayesiano supera consistentemente a regra da seguradora em termos de desempenho preditivo. Isto sugere que, embora a regra seja simples e interpretável, não captura todas as interações ou padrões presentes nos dados reais.

Seleção de Variáveis

Testaram-se diferentes combinações de atributos para avaliar a sua influência na performance do modelo:

- As quatro variáveis da regra dos 50 pontos (`Age`, `ChronicDiseases`, `MonthlySalary`, `Dependents`) obtiveram uma accuracy de aproximadamente **81%**.
- A adição de `Gender` e `PhysicalStatus` aumentou a accuracy para cerca de **83–84%**.
- O ganho adicional com a utilização das sete variáveis completas foi marginal, indicando que as quatro variáveis base já capturam grande parte da informação útil.

Contínuas vs Discretas

Foi também avaliado o impacto do tratamento das variáveis contínuas (`Age`, `MonthlySalary`) como tal versus a sua discretização em intervalos. O modelo com variáveis contínuas apresentou melhor desempenho (accuracy superior em cerca de 4 pontos percentuais), demonstrando que preservar a granularidade destas variáveis melhora a capacidade de previsão.

Teste de Normalidade

A suposição de normalidade foi testada com o teste de Shapiro-Wilk para `Age` e `MonthlySalary`. Os resultados revelaram que:

- A variável `Age` não segue uma distribuição normal para nenhuma das classes.
- A variável `MonthlySalary` aproxima-se mais da normalidade, sobretudo na classe "Aceita".

Estes resultados sugerem que a distribuição normal usada no modelo pode não ser a mais adequada em todos os casos, podendo justificar a utilização de abordagens mais flexíveis.

Abordagem com Misturas Gaussianas

Para ultrapassar estas limitações, foi explorada uma alternativa baseada em *Misturas Gaussianas* (*Gaussian Mixture Models*, GMM). Este tipo de modelação permite representar distribuições multimodais, ajustando múltiplas componentes gaussianas a cada classe da variável de saída (`Decision`).

A técnica foi aplicada às variáveis `Age` e `MonthlySalary`, ajustando duas e três componentes gaussianas para cada classe (Rejeita e Aceita). A análise visual dos contornos de densidade revelou que esta abordagem é capaz de capturar subestruturas relevantes nos dados, como a distinção entre subgrupos de indivíduos com diferentes perfis de idade e rendimento.

Fig. 3. Distribuição das GMM

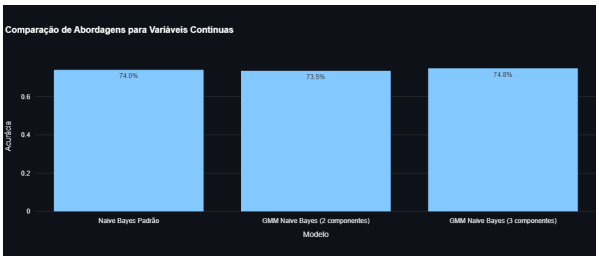

Fig. 4. Accuracy das GMM

Discussão

Os resultados obtidos evidenciam que o modelo Bayesiano proposto é eficaz na previsão da decisão de aquisição de seguro de vida, superando uma abordagem heurística baseada em regras fixas. Contudo, é importante refletir sobre as limitações da abordagem e as implicações dos resultados para uma eventual aplicação prática.

Relevância das Variáveis

A análise de seleção de atributos indicou que a maior parte da capacidade preditiva do modelo está concentrada em quatro variáveis: idade, doenças crónicas, salário e número de dependentes. Isto valida, em parte, a intuição da regra da seguradora, mas mostra também que outras variáveis — como género ou estado físico — contribuem marginalmente para o refinamento da decisão, embora com impacto limitado na accuracy final.

Tratamento de Variáveis Contínuas

O tratamento das variáveis `Age` e `MonthlySalary` como variáveis contínuas revelou-se benéfico para o desempenho do modelo. A discretização destas variáveis implicou perda de informação e redução na accuracy, o que é coerente com a natureza dos dados e a sua variabilidade. Apesar disso, a discretização pode ter vantagens interpretativas em contextos onde a transparência da decisão é mais relevante do que a exatidão.

Suposição de Normalidade

Embora o modelo base tenha inicialmente assumido distribuições normais para as variáveis contínuas, os testes de normalidade revelaram que essa suposição nem sempre é válida — especialmente no caso da idade, cujos valores não seguem uma distribuição normal em nenhuma das classes.

Para investigar alternativas, foi implementada e testada uma abordagem baseada em *Misturas Gaussianas* (GMM), permitindo modelar distribuições multimodais com maior flexibilidade. Esta abordagem demonstrou capacidade para capturar subestruturas nos dados, como subgrupos com diferentes perfis de idade e rendimento. Embora os ganhos em termos de accuracy tenham sido marginais, os resultados sugerem que o uso de misturas gaussianas pode ser benéfico em contextos mais complexos ou com maior variabilidade populacional.

Limitações do Modelo

- **Independência condicional:** A suposição de independência entre variáveis pode não refletir relações reais existentes (por exemplo, entre idade e estado de saúde).
- **Dependência dos dados rotulados:** O modelo aprende a partir de decisões humanas anteriores, que podem conter viés ou inconsistências.
- **Falta de validação externa:** A avaliação foi feita no mesmo conjunto de dados usado para treino (ainda que com filtros dinâmicos), o que poderá limitar a generalização dos resultados.

Conclusão

Este trabalho demonstrou a viabilidade de aplicar modelos probabilísticos, nomeadamente o classificador Naive Bayes, como ferramenta de apoio à decisão na contratação de seguros de vida. Através da utilização de dados históricos e de uma modelação baseada em inferência Bayesiana, foi possível prever com elevada accuracy a decisão de aquisição do seguro, superando significativamente a abordagem determinística tradicional da seguradora.

A análise evidenciou que a maior parte da capacidade preditiva reside em variáveis-chave como idade, doenças crónicas, salário e número de dependentes. No entanto, a inclusão de variáveis adicionais e o tratamento adequado das distribuições subjacentes — como ilustrado na abordagem com Misturas Gaussianas — permite refinar ainda mais as previsões, capturando padrões mais complexos.

A aplicação interativa desenvolvida torna estes resultados acessíveis a um público mais vasto, promovendo a transparência e facilitando a exploração dos dados e das recomendações. Para trabalhos futuros, sugere-se a integração de mais dados contextuais, a adoção de modelos mais robustos como árvores de decisão probabilísticas, e a análise da confiança nas previsões como fator de apoio à tomada de decisão ética e informada.

Em suma, este projeto ilustra como técnicas de fusão de informação podem contribuir de forma relevante para decisões sensíveis e personalizadas no setor dos seguros.

References

1. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29**(2-3), 103–130 (1997)
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009)
3. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC (2013)
4. Streamlit: The fastest way to build and share data apps. https://streamlit.io/ (2023)
5. Reynolds, D.A.: Gaussian mixture models. Encyclopedia of Biometrics **741**, 659–663 (2009)
6. Anderson, R.: The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford University Press (2007)
7. Frees, E.W., Derrig, R.A., Meyers, G.: Predictive Modeling Applications in Actuarial Science. Cambridge University Press (2014)