ContactLenses.csv foi o *dataset* escolhido para análise do trabalho final, com dados de prescrições de lentes de contacto (retirado de Cendrowska, 1987; os dados estão disponíveis através do Repositório de Aprendizagem Automática da UCI).

A variável a prever é se um doente deve ser colocado com lentes de contacto duras, moles ou sem lentes de contacto. Existem quatro atributos do paciente (todos dados categóricos em escalas nominais) que estão disponíveis para fazer essa análise:

• age {young, pre-presbyopic, presbyopic}

• spectacle-prescription {myope, hypermetrope}

• astigmatism {no, yes}

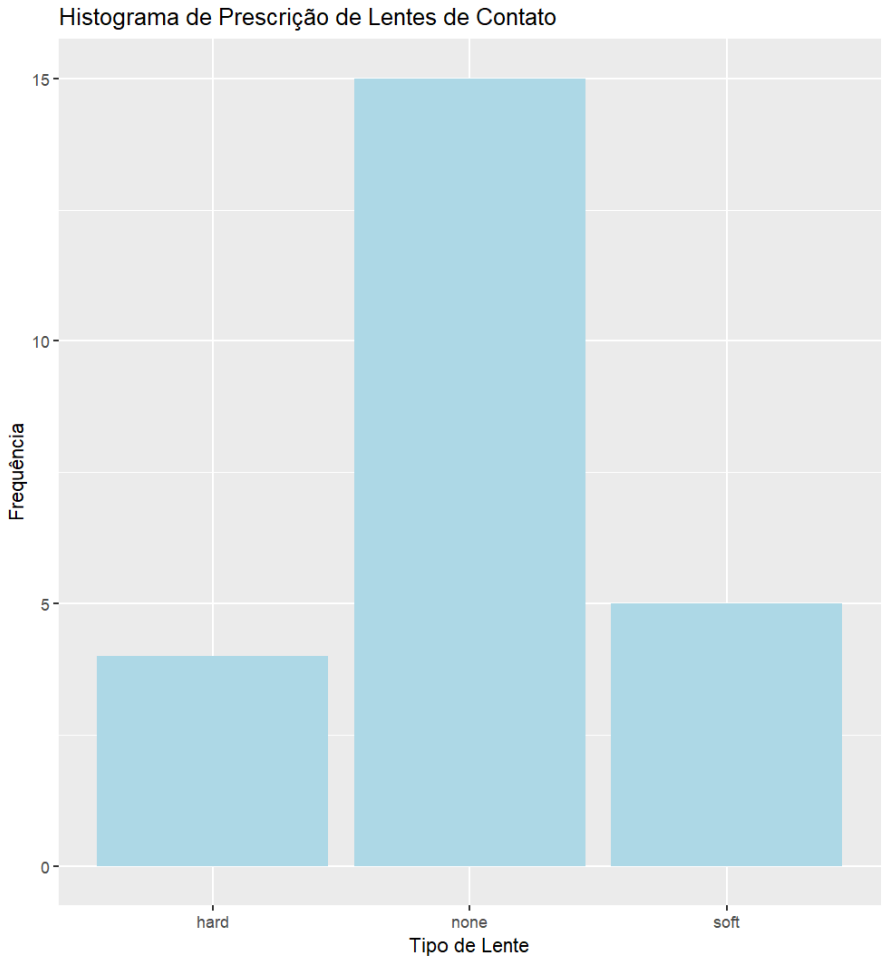
• tear-prod-rate {reduced, normal}

As dadas questões de Business Intelligence (BI) foram escolhidas com o objetivo de explorar os dados e contribuir também para as análises de Business Analytics (BA).

As primeiras duas questões de BI que surgiram ao analisar os dados foram:

1. **Qual a distribuição de pacientes que não devem usar lentes de contato em comparação aos que devem usar lentes?**

Essa pergunta busca entender a distribuição de recomendação de lentes com base nos dados.



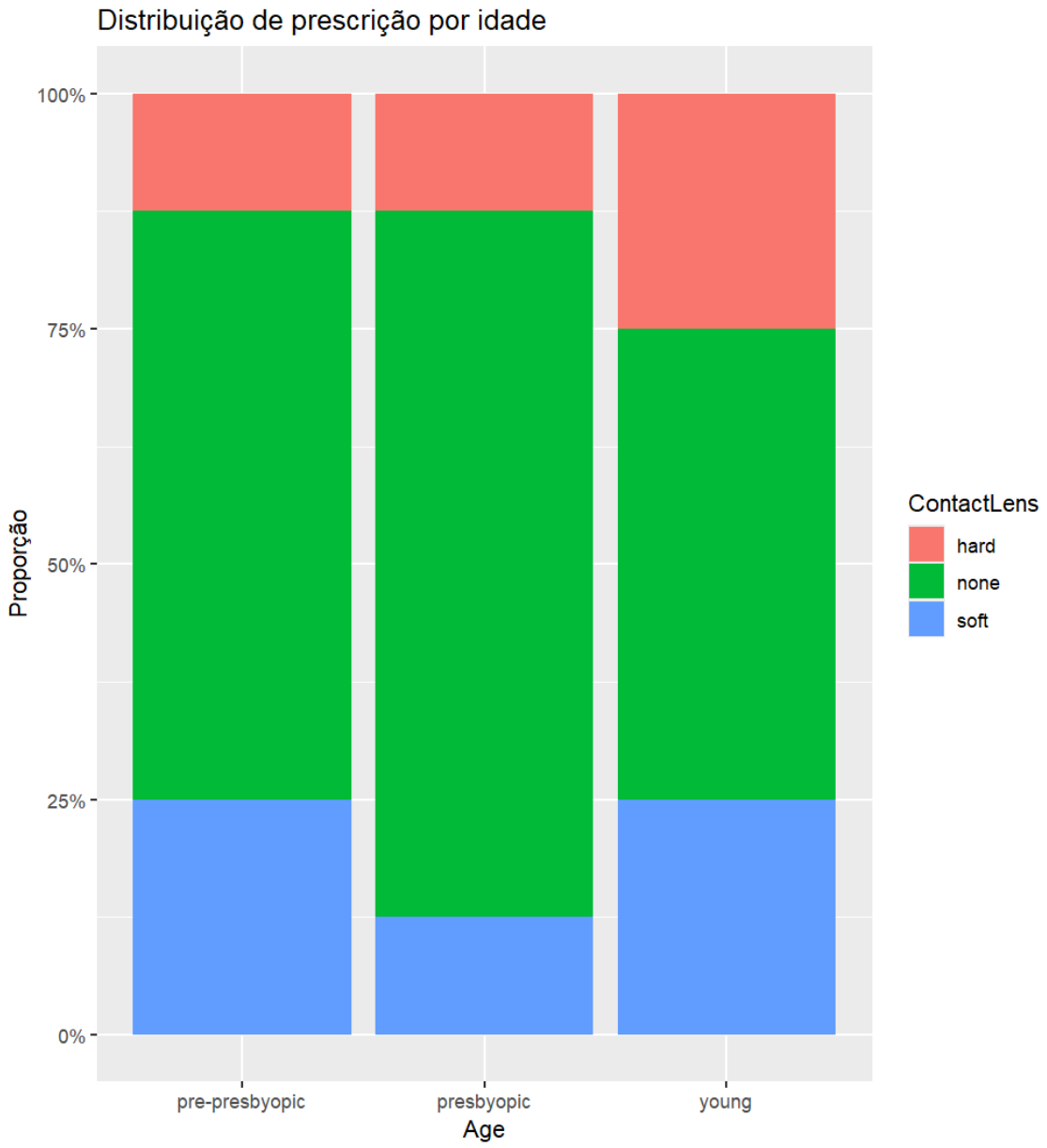
Esse histograma mostra a frequência de prescrição de diferentes tipos de lentes de contato, categorizadas em três tipos:

- hard (lentes rígidas)
- none (nenhuma prescrição)
- soft (lentes macias)

O tipo de prescrição mais comum é "none", com aproximadamente 15 ocorrências. As prescrições de lentes "hard" e "soft" têm frequências menores e semelhantes.

2. **Qual é a distribuição de tipos de lentes recomendadas por faixa etária (young, pre-presbyopic, presbyopic)?**

A ideia é comparar as tendências de prescrição de lentes de contato entre diferentes faixas etárias, mostrando como as necessidades e preferências variam com a idade



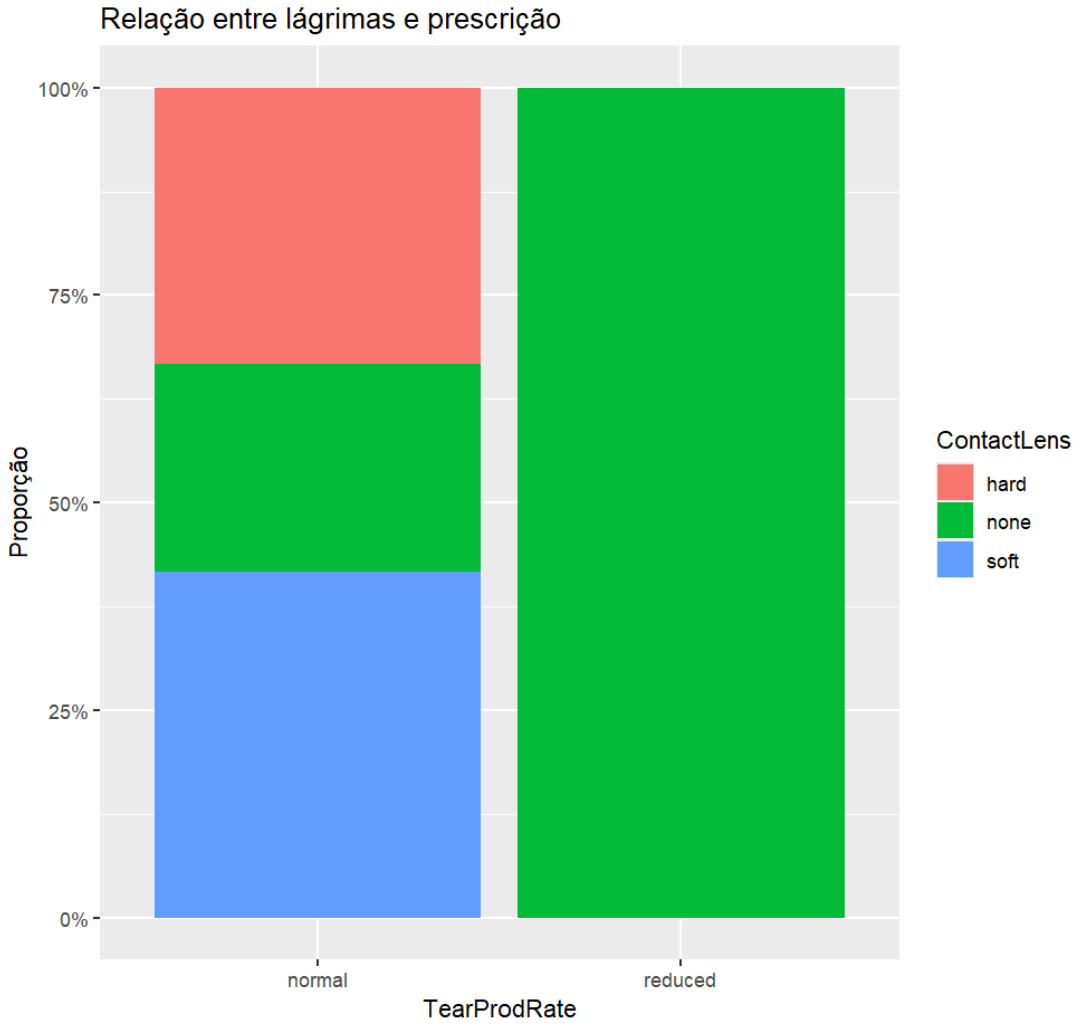
“Young”: Refere-se a pacientes jovens, tipicamente menores de 20 anos de idade, que não apresentam problemas visuais relacionados ao envelhecimento, como presbiopia.

“Pre-presbyopic”: Refere-se a pacientes que estão na faixa etária de 20 a 40 anos, ou seja, antes do surgimento da presbiopia. Esses pacientes ainda não têm perda da capacidade de focar objetos próximos relacionada ao envelhecimento.

“Presbyopic”: Refere-se a pacientes acima de 40 a 45 anos, que já apresentam presbiopia, uma condição natural de envelhecimento em que o olho perde a capacidade de focar em objetos próximos, geralmente exigindo o uso de óculos para leitura.

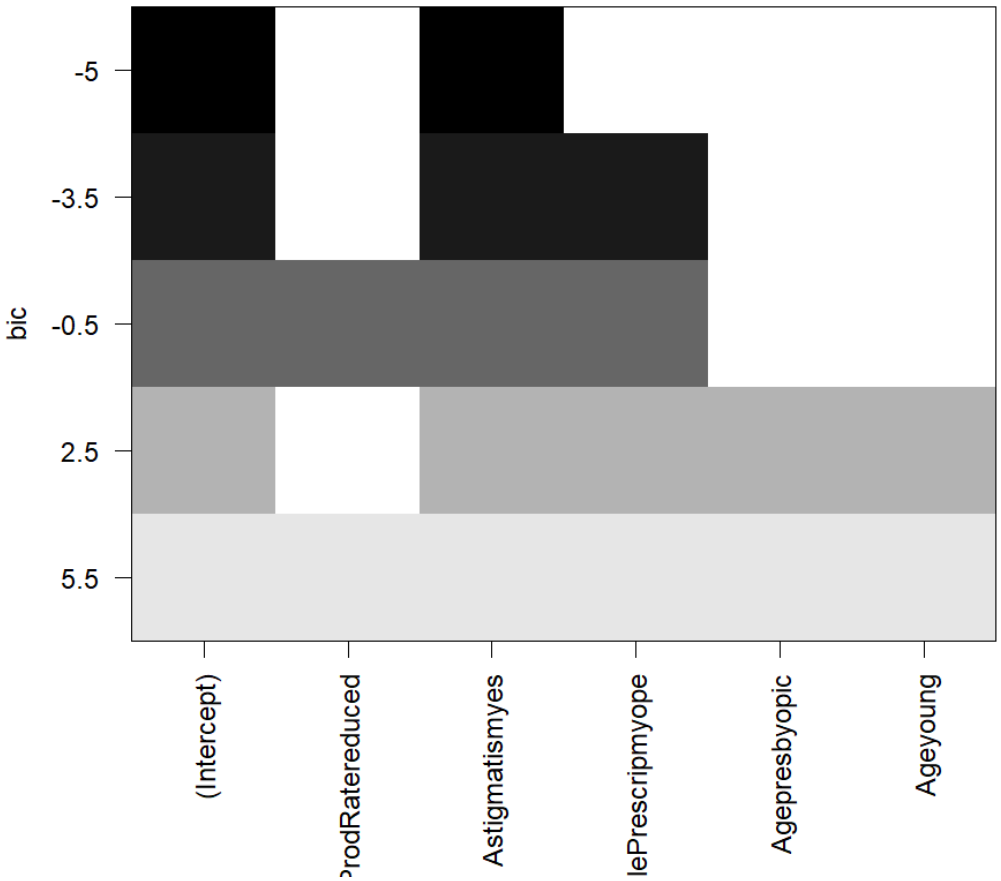
Os jovens têm a maior proporção de prescrições para lentes “hard” em comparação com os outros grupos. Os pré-presbíopes e jovens têm proporções similares de lentes “soft”, enquanto os presbíopes têm uma proporção menor. O grupo presbíope tem a maior proporção de pessoas que não usam lentes de contato. As lentes macias são mais comuns que as duras em todos os grupos etários.

3. **Qual a relação entre a produção de lágrimas dos pacientes e a prescrição de lentes?**



Pessoas com produção normal de lágrimas têm muito mais probabilidade de usar lentes de contato do que aquelas com produção reduzida. A produção reduzida de lágrimas parece ser um fator limitante significativo para a capacidade de usar lentes de contato, sugerindo que a lubrificação dos olhos, é crucial na determinação das opções de correção visual disponíveis para um paciente.

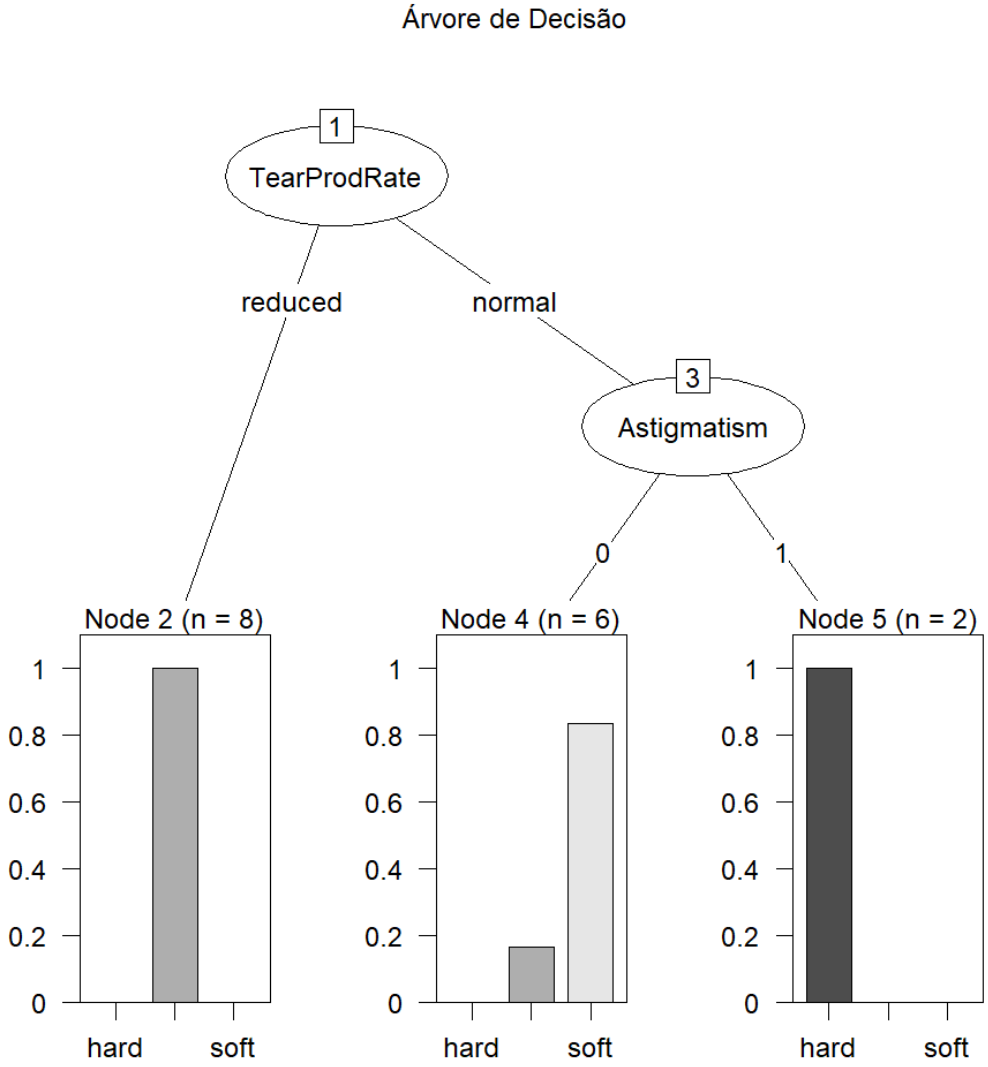
Após exploração dos dados e variáveis, plotamos um gráfico que indica quais variáveis são mais importantes para o ajuste do modelo e quais têm menor influência, com base no BIC (Bayesian Information Criterion).



Valores **negativos** ("TearProdRatereduced" e "Astigmatismyes") indicam que essas variáveis têm um **impacto mais forte** e melhoram o ajuste do modelo.

Valores **positivos** ("Agepresbyopic" e "Ageyoung") sugerem que essas variáveis são **menos importantes** ou que o modelo penalizou essas variáveis por aumentar a complexidade sem melhorar muito o ajuste.

Após um sample dos dados para garantir uma distribuição imparcial, foi feita a divisão da amostra em teste e treino e criamos nossa árvore de decisão, usando apenas o conjunto de treino. O modelo foi treinado para prever o tipo de prescrição de lente de contato com base nas outras variáveis do conjunto de dados.



1 – “TearProdRate”:

O primeiro fator considerado é a taxa de produção de lágrimas.

Há duas possibilidades: "reduced" ou "normal".

2 – “Reduced”:

Se a produção de lágrimas é reduzida, a decisão nem é “hard”, nem “soft”.

Total de 8 casos (n = 8).

“Normal”:

Se a produção de lágrimas é normal, a árvore considera um segundo fator: o astigmatismo.

3 – “Astigmatism”:

Para pacientes com produção normal de lágrimas, o grau de astigmatismo influencia a decisão. Divide-se em dois grupos: astigmatismo 0 ou 1 que no caso significa respectivamente “no” e “yes”.

4 – “Astigmatism” 0 (no):

Preferência por lentes "soft".

Total de 6 casos (n = 6).

5 – “Astigmatismo” 1 (no):

Há uma preferência por lentes "hard".

Total de 2 casos (n = 2).

Questões de Business Analytics (BA):

1. **Qual é a probabilidade de um paciente jovem com astigmatismo e produção normal de lágrimas receber a recomendação de lentes rígidas em vez de macias ou nenhuma?**

No node 4, que corresponde a pacientes com produção normal de lágrimas e astigmatismo, a distribuição é:

Node 5 (n = 2), onde:

- A probabilidade de lentes rígidas (hard) é 2 em 2
- A probabilidade de lentes macias (soft) é 0 em 2

Ou seja, um paciente jovem com produção normal de lágrimas e astigmatismo tem 100% de probabilidade de receber a recomendação de lentes rígidas e 0% de probabilidade de receber a recomendação de lentes macias.

2. **Quais combinações de atributos têm maior probabilidade de levar a erros nas previsões sobre o tipo de lente que um paciente deve usar, considerando a produção de lágrimas e a presença de astigmatismo?**

Em um contexto de previsão, erros de classificação são possíveis porque o modelo está tentando prever um resultado com base em dados disponíveis. Quando a proporção de classes é muito próxima, a classificação se torna mais propensa a erros.

Node 2 (“TearProdRate”: “reduced”, n = 8):

- “soft”: 100%
- “hard”: 0%

Node 4 (“TearProdRate”: “normal”, “Astigmatism”: 0, n = 6):

- “soft”: 83%
- “hard”: 17%

Node 5 (“TearProdRate”: “normal”, “Astigmatism”: 1, n = 2):

- “hard”: 100%
- “soft”: 0%

Maior risco de erro:

Node 4 (produção de lágrimas normal + astigmatismo presente): É um node de maior risco, pois tem uma distribuição menos definida (17% hard e 83% soft).

Node 5 (produção de lágrimas normal + sem astigmatismo): Aqui a distribuição é mais balanceada (67% hard e 33% soft). Um modelo o poderia errar ao predizer majoritariamente hard e não capturar a variabilidade presente. Esse é um node de risco significativo, pois é o mais balanceado entre as classes.

As combinações de atributos que têm maior chance de gerar erros são aquelas que caem nos nodes 4 e 5, que possuem distribuições de classes mais balanceadas e complexas de capturar.