**Questão 1**

* 1. Escolhi o algoritmo kmeans, o algoritmo apresenta bons desempenhos em vários trabalhos relacionados na área de cluster e também é utilizado como benchmark em vários artigos, os dados dos problemas apresentam em suas maiorias variáveis dicotômicas e algoritmo de distância utilizado é a euclidiana que se encaixa muito bem nesses tipos de variáveis.

**1.2-**Eu removi variáveis que possuíam classe com valores baixos de 30 observações em uma das classes binárias, isto afetada toda inferência de aproximação quando temos valores abaixo com isso gerando cluster a mais com baixos valores de composição. Outro ponto relevante é a construções de variáveis a partir da tabela “heroes\_information”, percebi que existia heróis duplicados, com isso fiz uma transformação de grão, a nova de dados ficou com os heróis únicos e para as variáveis que ele tinha mais informações diferentes, calculei a variável com o percentual das classes que ele tinha, por exemplo: O herói Spider-Man tem três informações e a variável “Eye color” tinha 3 valores distintos “hazel”, “red” e ” brown”, com a variável transformada ficou “Eye color$hazel” com o valor de 0,333 , “Eye color$red” com o valor de 0,333 e “Eye color$brown” com o valor de 0,333. Para as variáveis numéricas criei faixas de acordo com o quantil em 5 faixas, foram elas: “Height” e “Weight”.

**Questão 2**

Para selecionar o numero de cluster utilizei a técnica “Elbow” que é uma técnica voltada para identificar mudança em média dos valores em todos os cluster a medida que aumenta a quantidade, quando aumentamos os clusters se verificar se houve uma diferença abaixo de 200, quando ocorre chamo esse ponto de “cutuvelo” , baseado nesses parâmetro a quantidade indicada foi de 4 clusters.

Analisando o resultado dos clusters, a variável “Longevity”, temos as médias dos clusters abaixo:

|  |  |
| --- | --- |
| Clusters | Longevity |
| C1 | 44,58% |
| C2 | 7,62% |
| C3 | 10,57% |
| C4 | 5,15% |

O cluster 1 possui uma quantidade maior de heróis com longevidade em comparação ao restantes dos clusters e o cluster 4 possui o menor percentual de heróis com longevidade.

**Questão 3**

**3.1 -** Um classificador Naive Bayes é um classificador probabilístico simples baseado na aplicação de teorema de Bayes com fortes hipóteses de independência entre seus atributos, em termos simples, um classificador de Bayes assume que a presença ou ausência de uma característica particular, não está relacionado com a presença ou ausência de qualquer outro elemento, tendo em conta a classe variável.

Por exemplo, um herói pode ser considerado como bom, se é forte, com longevidade, e com uma altura de 3m. Um classificador Naive Bayes considera cada um desses recursos para contribuir de forma independente para a probabilidade de que este herói seja forte, independentemente da presença ou ausência de outras características.

**3.2 -** Para rodar o modelo foram realizados pré-processamento em cima dos dados, para as variáveis dependente numéricas com missing foi adicionado o valor mínimo no missing e para as variáveis categóricas com missing foi adicionada a categoria “False”, com isso, depois da primeira rodagem é analisada as probabilidades condicionais das variáveis explicativas para cada classe do alvo, se a probabilidade de separação das classes do alvo não for significativa essa variável explicativa tem pouco poder de separação das classes assim ela pode ser removida no processo de aprendizagem, segue um exemplo na tabela abaixo:

|  |  |  |
| --- | --- | --- |
| Y | The Force | |
| False | True |
| bad | 98,10% | 1,90% |
| good | 99,73% | 0,27% |
| neutral | 100,00% | 0,00% |

A variável “The force” ilustra bem apresentando pouca variação entre as probabilidades da classe “False” contra as três classes do alvo, com isso trazendo pouco poder de predição para o modelo.

**3.3 -** Eu avalio os resultados em cima de uma base de teste que foi separada de forma aleatória, minha escolha foi de 25% para o teste e o complementar para o treinamento, em seguida calculo a taxa de acerto geral e por classe do alvo. Para o experimento que rodei o acerto geral foi de 67,28% e para cada classe segue a tabela abaixo:

|  |  |
| --- | --- |
| Classes | Acerto |
| bad | 49,94% |
| good | 76,86% |
| neutral | 0% |

**Questão 4**

**4.1**-Escolho a técnica GBM (Gradient Boosting Machine) este técnica de ensemble de classificadores fracos como por exemplo uma arvore de decisão que se encaixa muito bem por questão da aplicação de suas segmentações das variáveis, para esse problema pelo fato de ter muitas variáveis dicotômicas que apresentam valores baixos em uma de suas classes a arvore de decisão extrai bem essa informação com suas divisões e expansão dos seus nós, com isso gerando uma performance superior a muitas técnicas existentes como exemplo uma rede neural, Naive Bayes e entre outras.

**4.2**-A comparação será realizada em cima da mesma base de teste que foi separada já na questão anterior, o acerto geral e por classe será comparado e quem apresentar o melhor resultado será o escolhido como a técnica que encaixa com melhor desempenho, segue a tabela abaixo com os valores de comparação:

|  |  |
| --- | --- |
| Modelos | Acerto Geral |
| Bayes | 67,28% |
| GBM | 70,33% |

|  |  |  |
| --- | --- | --- |
| Classes | Acerto Bayes | Acerto GBM |
| bad | 49,94% | 53,06% |
| good | 76,74% | 79,07% |
| neutral | 0,00% | 0,00% |

O modelo GBM apresentou melhores resultados em comparação ao modelo de Bayes.

**Questão 5**

**Beyond Good and Evil**

**5.1**-Irei utilizar duas técnicas uma é o SVM (Support Vector Machine) e a outra é a GBM, o SVM é uma técnica que apresenta bons resultados no quesito em classificação e regressão, o ajuste que a técnica realiza para encontrar seu melhor suporte o torna um modelo robusto e suscetível a overfitting e underfitting, o GBM como foi mencionado na questão anterior se encaixa bem ao problema.

**5.2-**A técnica para validar os resultados das técnicas será o RMSE (Root-Mean Square Deviation) representa o desvio padrão da amostra das diferenças entre os valores previstos e os valores observados.

Segue abaixo as tabelas de comparação entre as duas técnicas escolhidas para este problema de regressão:

|  |  |
| --- | --- |
| Modelos | RMSE |
| GBM | 80,85 |
| SVM | 95,65 |

O GBM apresentou um resultado superior, com isso ele é o escolhido novamente.

**Questão 6 - Analysis**

O problema apresentado na base de dados para clustering foi o trazer a informação extra da tabela “heroes\_information” para a “super\_heroes” sabendo que para realizar clustering a granularidade tem que ser única e a tabela “Heroes\_information” trazia mais de uma informação para a “super\_heroes” a solução viável foi a transformação de granularidade com isso a tabela “heroes\_information” ficou com a granularidade única e amplificando o numero de variáveis.

Para o problema de classificação o obstáculo foi a falta de informação de alguns heróis quando se realiza o left join entre a tabela “super\_heroes” e “heroes\_information”, a solução para contornar foi preenchimento de missing value, para as variáveis numéricas com o mínimo, máximo, média ou mediana e a categórica chamando de outra classe ou uma classe de negação se for dicotômica.

Para a regressão foi falta de informação na variável alvo que representava 20% dos dados, a solução para este caso foi remover na base de treinamento esses registros e depois gerar a predição na base de teste desses registros com valores ausentes no alvo, para realizar a medida de desempenho no teste esses registros foram removidos.

**Questão 6 - Bonus**

Apliquei a técnica CHAID (Chi-square Automatic Interaction Detector) que é uma ferramenta usada para descobrir a relação entre variáveis, onde constrói um modelo preditivo, ou árvore, para ajudar a determinar como as variáveis ​​se fundem melhor para explicar o resultado na variável dependente dada. Na análise, dados nominais, ordinais e contínuos podem ser usados, onde preditores contínuos são divididos em categorias com número de observações aproximadamente igual.

Foram extraídas duas informações importantes para a classe “good”, são elas:

Regra 1: Super.Strength %in% c(\"False\") & Weapon.based.Powers %in% c(\"False\")"

Regra 2: "Super.Strength %in% c(\"True\") & Weight %in% c(\"[2,54)\", \"[54,77)\")"

A Regra 1 representa 46,77% da base dados e com uma taxa de ”good” de 77,65% e a Regra 2 representa 11,55% e a taxa de “good” de 77,38%, ou seja, temos dois cubos de regra que representa mais de 50% da base com mais de 77% de heróis “good”.