U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

SHOULD WE LOAN?

PROJETO DATA MINING I

MARIA PAIS, UP202308322
MÓNICA ARAÚJO, UP202005209

INDICE

DEFINIÇÃO DO PROBLEMA

Um banco planeja melhorar a qualidade do seu serviço ao cliente. Um desafio particular que enfrenta é a ambiguidade em torno da identificação de bons clientes e maus clientes.

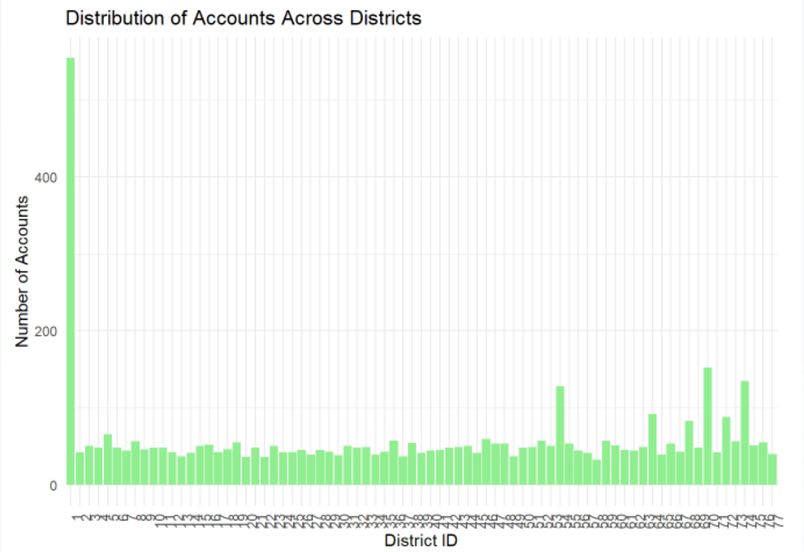
Este projeto visa utilizar técnicas de extração de dados neste conjunto de dados para ajudar os gestores do banco a compreender melhor os seus clientes e a identificar se um empréstimo será concluído com êxito.

DATA UNDERSTANDING

Data Understanding é uma fase crucial do processo de análise de dados. Envolve a obtenção de conhecimentos sobre a natureza, estrutura, conteúdo e qualidade de um conjunto de dados. O objetivo da compreensão dos dados é familiarizar-se com os dados, identificar potenciais problemas e lançar as bases para as fases subsequentes da análise de dados.

Os principias pontos retirados:

- os dados observados vão desde 1993 até 1997;
- só temos informação sobre os loans desde 1993 até 1996;
- os distritos com mais contas são o nº1, nº74 e nº70;
- a região com mais habitantes é Prague
- cada account tem 1 ou 0 loans
- existem 282 loans com status “1” e 46 com status “-1”



DATA PREPARATION

01

CONVERSÃO DO TIPO DE DADOS

Garantia de consistencia e adequação dos tipos no

Ex: conversão das datas para formatos adequados

02

TRATAMENTO DE MISSING VALUES

Remoçao de Propriedades com mais de 50% de NA

Imputação dos missing values

03

REMOÇÃO DE DUPLICADOS

Indentificação de valores duplicados: por linha e por algumas linha de conjunto de coluna(s)

04

ENCODING DE VARIÁVEIS CATEGÓRICAS

Transformação de variaveis em formato adequado através do metodo One-hot encoding

DATA PREPARATION

01

FEATURE ENGINEERING

Criação de novas características relevantes

Exemplo: Calculo do dinheiro na conta antes de um emprestimo

02

DIVISÃO DOS DADOS

Separação dos dados em conjuntos de treino e teste

03

PADRONIZAÇÃO DAS FEATURES

Aplicação da padronization utilizando o metodo z-scale normalization em variaveis selecionadas

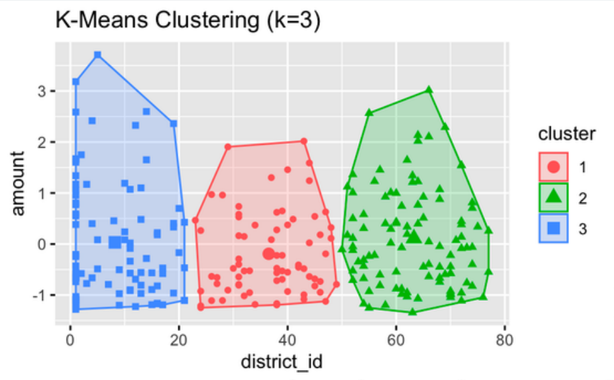
04

ARMAZENAMENTO OS DADOS PROCESSADOS

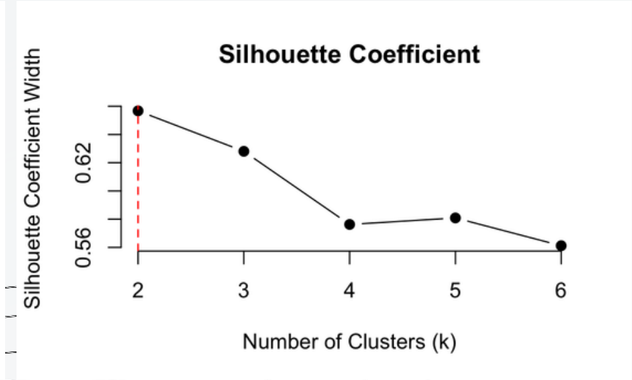
Armazenamento dos dados em ficheiros RData, prontos para serem usados em passos posteriores

DESCRIPTIVE MODELING

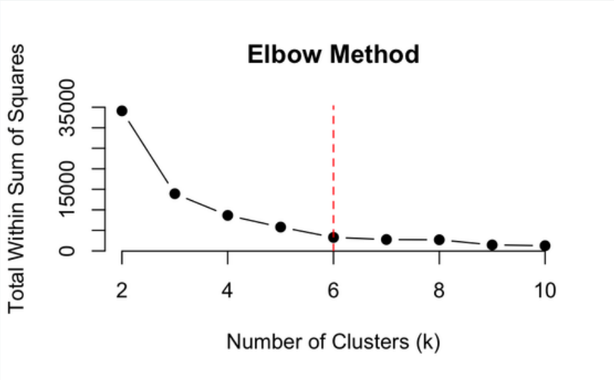
K-MEANS



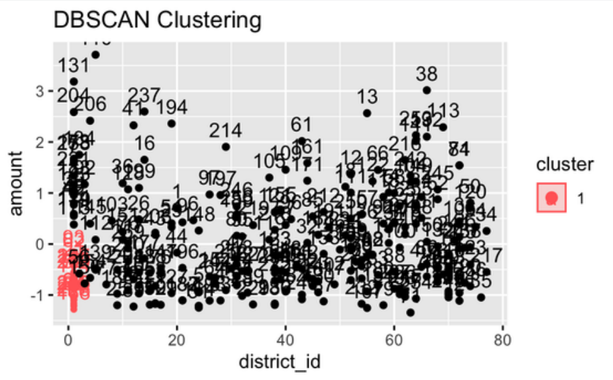
K-MEAN - SILHOUETTE METHOD

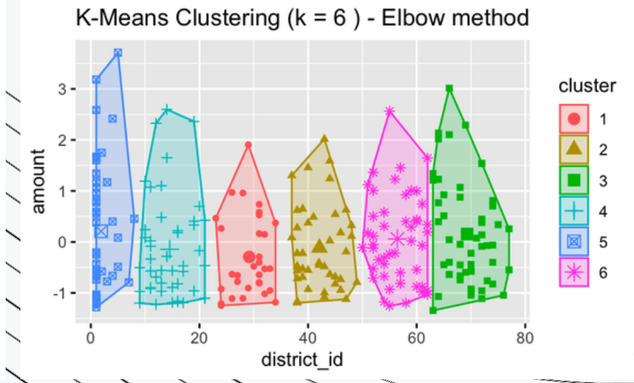


K-MEAN - ELBOW METHOD



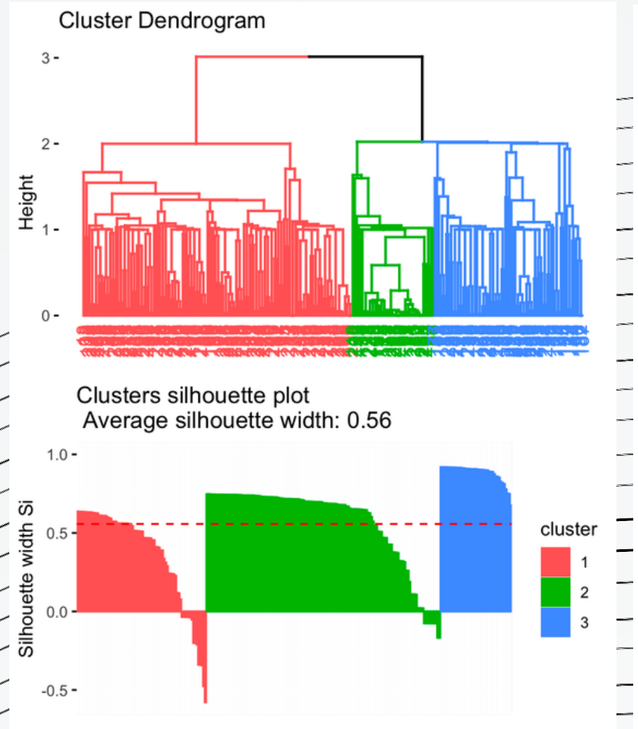
DBSCAN



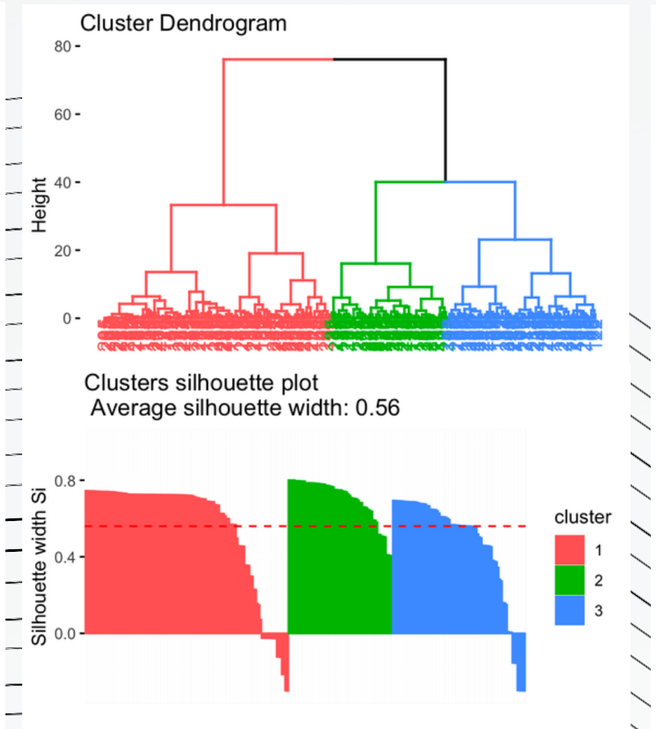


DESCRIPTIVE MODELING

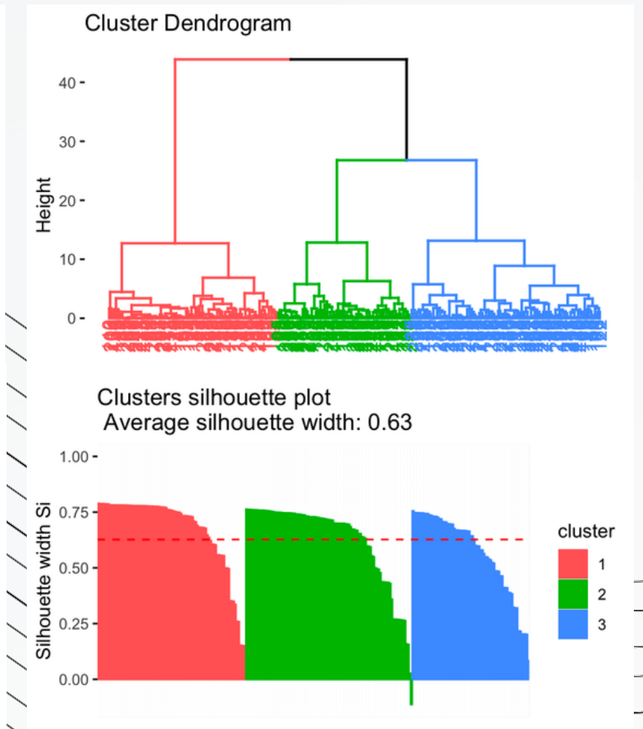
HIERARCHICAL - AGGLOMERATIVE METHOD - SINGLE LINK



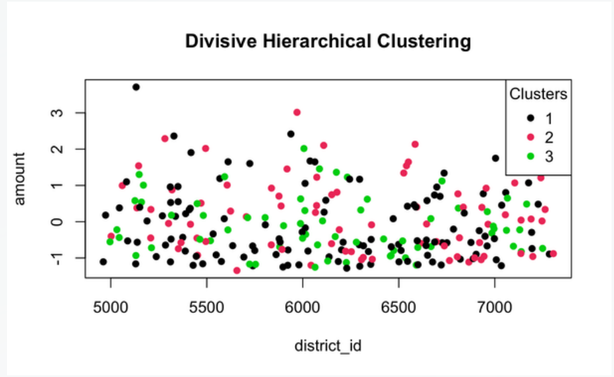
HIERARCHICAL - AGGLOMERATIVE METHOD - COMPLETE LINK



HIERARCHICAL - AGGLOMERATIVE METHOD - COMPLETE LINK



DIVISIVE METHOD



PREDICTIVE MODELING

São um tipo de algoritmo de aprendizagem supervisionada que imita a estrutura de uma árvore para tomar decisões. A estrutura em árvore é constituída por nós, em que cada nó representa uma decisão baseada numa determinada caraterística, e os ramos representam os resultados possíveis dessa decisão

DECISION TREES

Modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. O objetivo da regressão linear é encontrar a melhor relação linear (uma linha reta no caso da regressão linear simples) que preveja a variável dependente com base nas variáveis independentes.

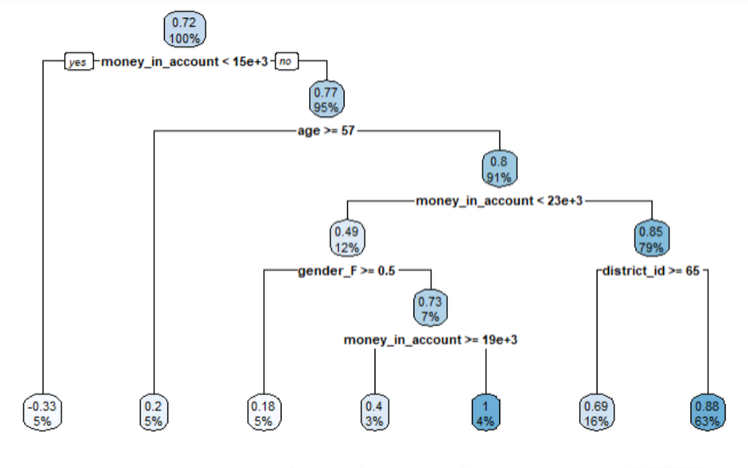
LINEAR REGRESSION

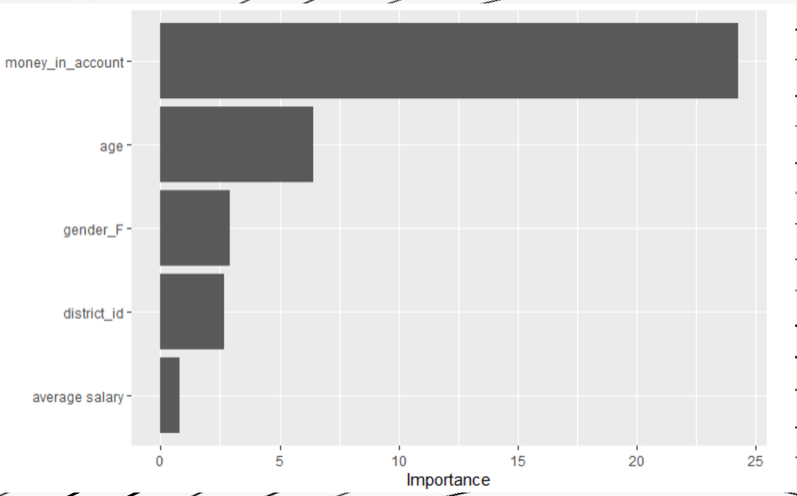
O Random Forest é uma técnica de aprendizagem de conjunto que combina as previsões de vários modelos individuais para melhorar o desempenho global e a generalização. O Random Forest é particularmente poderoso e versátil, sendo frequentemente utilizado para tarefas de classificação e regressão.

RANDOM FOREST

PREDICTIVE MODELING

DECISION TREES





LINEAR REGRESSION

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.1228  0.1190  0.2671  0.3385  0.4368

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.7627712  0.0958909   7.955 5.81e-14 ***
age              -0.0188916  0.0429277  -0.440  0.66025
district_id      -0.0005803  0.0019974  -0.291  0.77166
gender_F         -0.0343221  0.0856065  -0.401  0.68881
money_in_account  0.1132278  0.0427120   2.651  0.00853 **
`average salary`  0.0061305  0.0492193   0.125  0.90097
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6863 on 256 degrees of freedom
Multiple R-squared:  0.02915,   Adjusted R-squared:  0.01019
F-statistic: 1.537 on 5 and 256 DF,  p-value: 0.1786
```

RANDOM FOREST

```
predictions -1  1
         -1  0  0
          1 10 56
[1] 0.8484848
```

O R-squared do modelo de regressão linear é bastante baixo, indicando que o modelo explica apenas uma pequena proporção da variação na variável de resposta.

O modelo de random forest tem maior exatidão, precisão e recuperação, o que sugere um melhor desempenho global de classificação em comparação com o modelo de regressão linear.

O modelo escolhido é o random forest

CONCLUSÃO E TRABALHO FUTURO

Como modelo final foi escolhido random forest

Como trabalho futuro fica a sugestão de implementação de redes neuronais e o teste de mais modelos.

OBRIGADA PELA ATENÇÃO