Exercicio Extra lista 4

Davi Wentrick Feijó - 200016806

2023-05-17

Sobre o Artigo

O artigo fala sobre a Mineração de Dados Educacionais (EDM) que é uma área que busca processar e interpretar dados educacionais para obter conhecimento útil. Um dos principais objetivos da EDM é prever o desempenho dos alunos. No entanto, a alta dimensionalidade dos conjuntos de dados educacionais pode dificultar a análise e a precisão das previsões (Cruse of Dimensionality). Portanto, este estudo propõe o uso da Análise de Componentes Principais (PCA) como uma técnica para reduzir a dimensionalidade dos dados e extrair conhecimento relevante. O PCA é uma técnica bem conhecida que captura a variabilidade dos dados usando poucas dimensões. O estudo também utiliza dois modelos de classificação populares, Máquinas de Vetores de Suporte e Naive Bayes, para realizar as previsões de desempenho dos alunos. Os resultados experimentais mostraram a eficácia do método proposto. Este estudo contribui para o uso do PCA na análise de dados educacionais e na redução da dimensionalidade para tarefas de previsão de desempenho dos alunos.

Sobre o banco utilizado

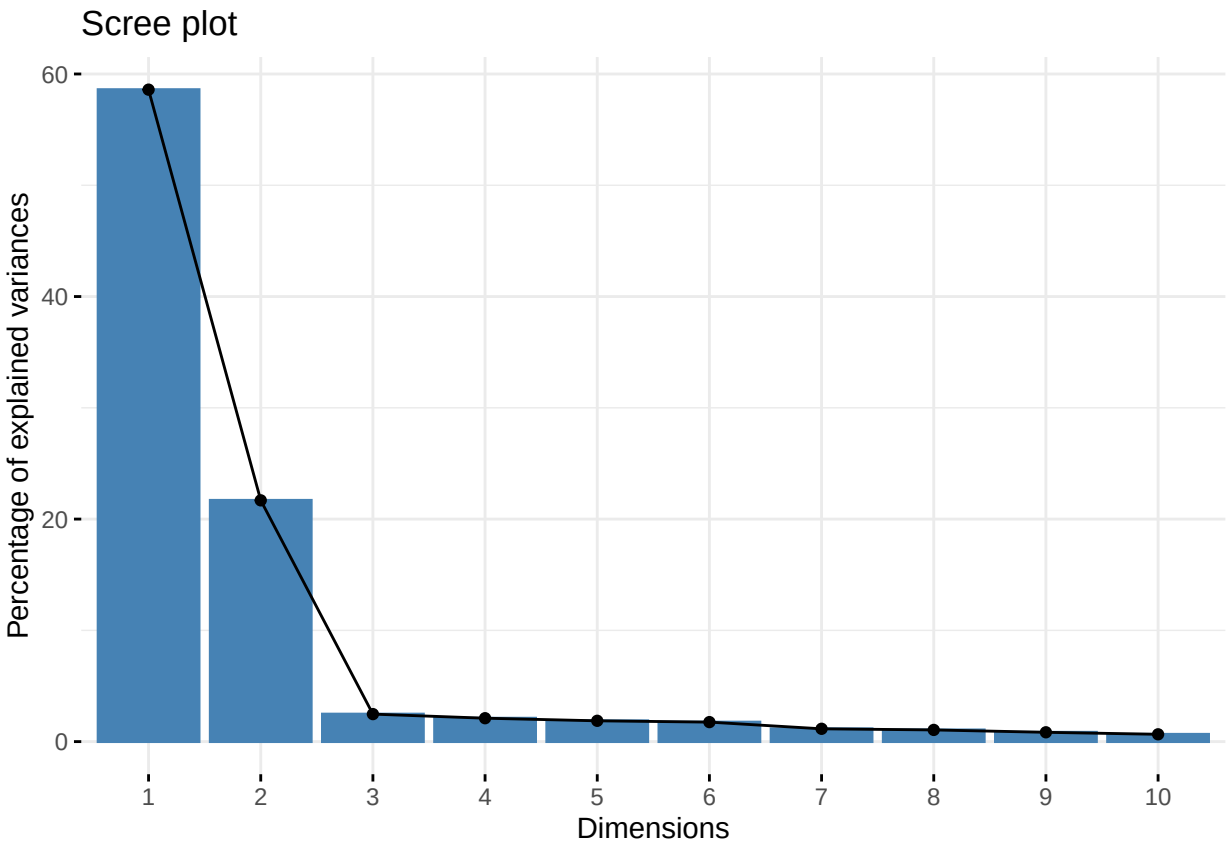
Os conjuntos de dados descrevem as realizações dos alunos no ensino médio de duas escolas portuguesas em relação a duas disciplinas: Português e Matemática. Os atributos dos dados incluem notas dos alunos, seus registros demográficos, sociais, financeiros, pessoais e acadêmicos. Esses dados foram coletados por meio de relatórios escolares e questionários. O primeiro conjunto de dados (Dataset I) contém 649 alunos da disciplina de Português e é descrito por 33 atributos. O segundo conjunto de dados (Dataset II) é caracterizado pelos mesmos atributos e se refere às realizações finais dos alunos na disciplina de Matemática.

A nota final é considerada como o atributo de classe, uma vez que estamos interessados em prever o desempenho dos alunos. Em seus estudos, Cortez e Silva relataram que o atributo de classe G3 apresenta alta correlação com os atributos G2 e G1. Isso ocorre porque G3 é a nota do último ano (emitida no 3º período), enquanto G1 e G2 correspondem às notas do primeiro e segundo períodos.

Tratamento de dados para aplicação da PCA

O objetivo desta etapa é preparar os dados brutos em uma forma adequada que permita a aplicação das técnicas de aprendizado de máquina. Uma análise preliminar nas instâncias de dados e nos atributos dos Conjuntos de Dados I e II mostrou que é necessária alguma forma de pré-processamento, uma vez que os atributos são de diferentes tipos (binários, numéricos e nominais). Primeiro, as instâncias de dados com valores ausentes são removidas, para que possamos lidar com dados consistentes. Em seguida, como existem atributos nominais, cada um deles é transformado em variáveis dummy, que podem ser definidas como atributos binários que podem assumir o valor “0” ou “1” para indicar a ausência ou presença de um valor categórico específico.

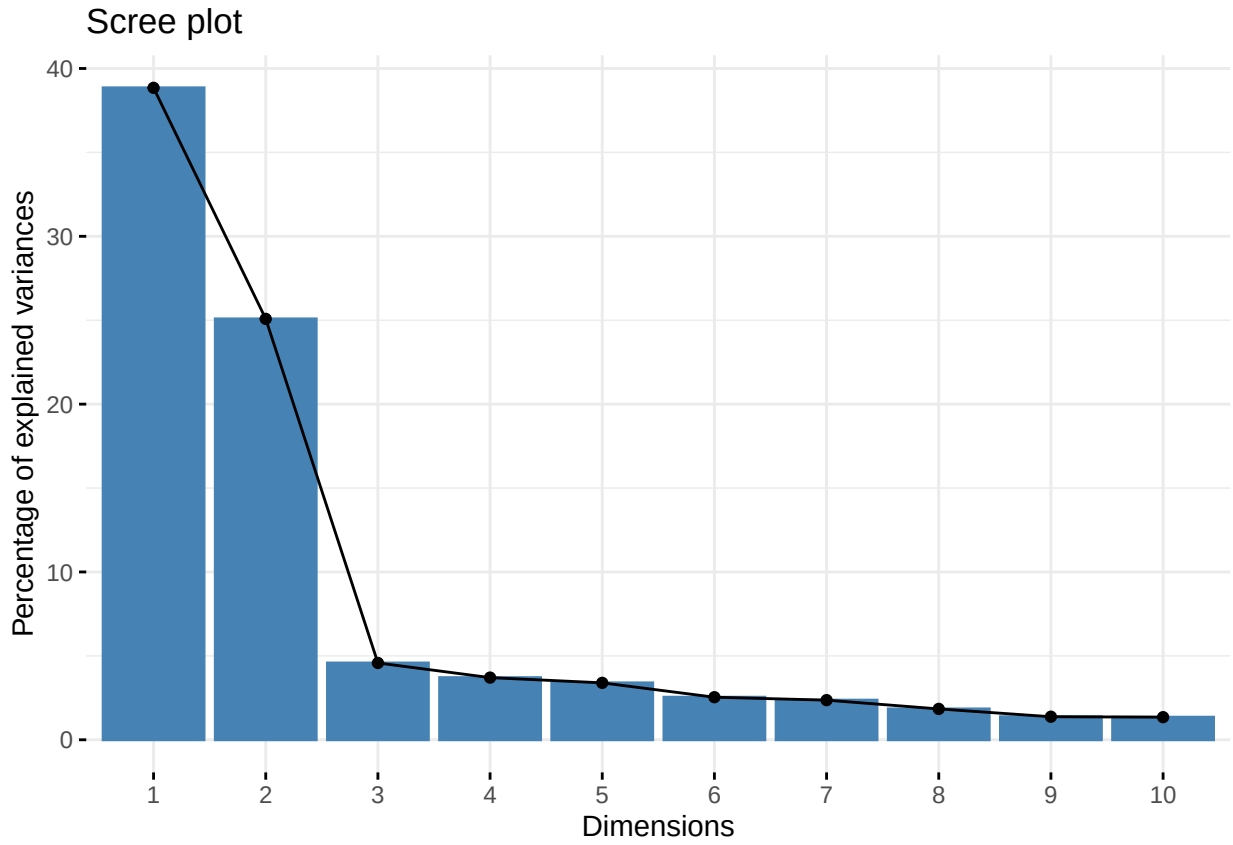
Analise das notas de Mátematica




```
##  [1] 58.592 21.684  2.473  2.104  1.871  1.753  1.152  1.052  0.838  0.658
```

```
## # A tibble: 6 x 33
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob    reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>    <chr>   <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home  teacher course
## 2 GP     F        17 U       GT3     T           1     1 at_home  other   course
## 3 GP     F        15 U       LE3     T           1     1 at_home  other   other
## 4 GP     F        15 U       GT3     T           4     2 health   servic~ home
## 5 GP     F        16 U       GT3     T           3     3 other    other   home
## 6 GP     M        16 U       LE3     T           4     3 services other   reput~
## # i 22 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
## #   failures <dbl>, schoolsup <chr>, famsup <chr>, paid <chr>,
## #   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
## #   romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>, Dalc <dbl>,
## #   Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

Analise das notas de Portugues




```
##  [1] 38.848 25.080  4.578  3.705  3.393  2.537  2.362  1.840  1.374  1.348
```

```
## # A tibble: 6 x 33
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob    reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>    <chr>   <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home  teacher course
## 2 GP     F        17 U       GT3     T           1     1 at_home  other   course
## 3 GP     F        15 U       LE3     T           1     1 at_home  other   other
## 4 GP     F        15 U       GT3     T           4     2 health   servic~ home
## 5 GP     F        16 U       GT3     T           3     3 other    other   home
## 6 GP     M        16 U       LE3     T           4     3 services other   reput~
## # i 22 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
## #   failures <dbl>, schoolsup <chr>, famsup <chr>, paid <chr>,
## #   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
## #   romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>, Dalc <dbl>,
## #   Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

Dentro desse artigo foi usado o scree plot para ter um representaçaõ grafica da variancia de cada componente, uma alternativa pode ser mostrar a probabilidade cumulativa. Porém para esse tipo de gráfico nao tem muitas alternativas de representação.