Icon

Description automatically generated

Universidade do Minho

Departamento de Informática

Sistemas Baseados em Similaridade

Trabalho Prático Individual 1

Gonçalo Almeida (A84610)

3 de Novembro de 2020

**T1.** Carregar, no Knime, o dataset descarregado. Aplicar nodos para exploração de dados, i.e., analisar os dados em relação à sua:

a. Tendência central;

b. Dispersão estatística;

c. Correlação entre features.

Diagram

Description automatically generated

Figure - Exploração de dados

Table

Description automatically generated

Figure 2 -View do nodo Data Explorer

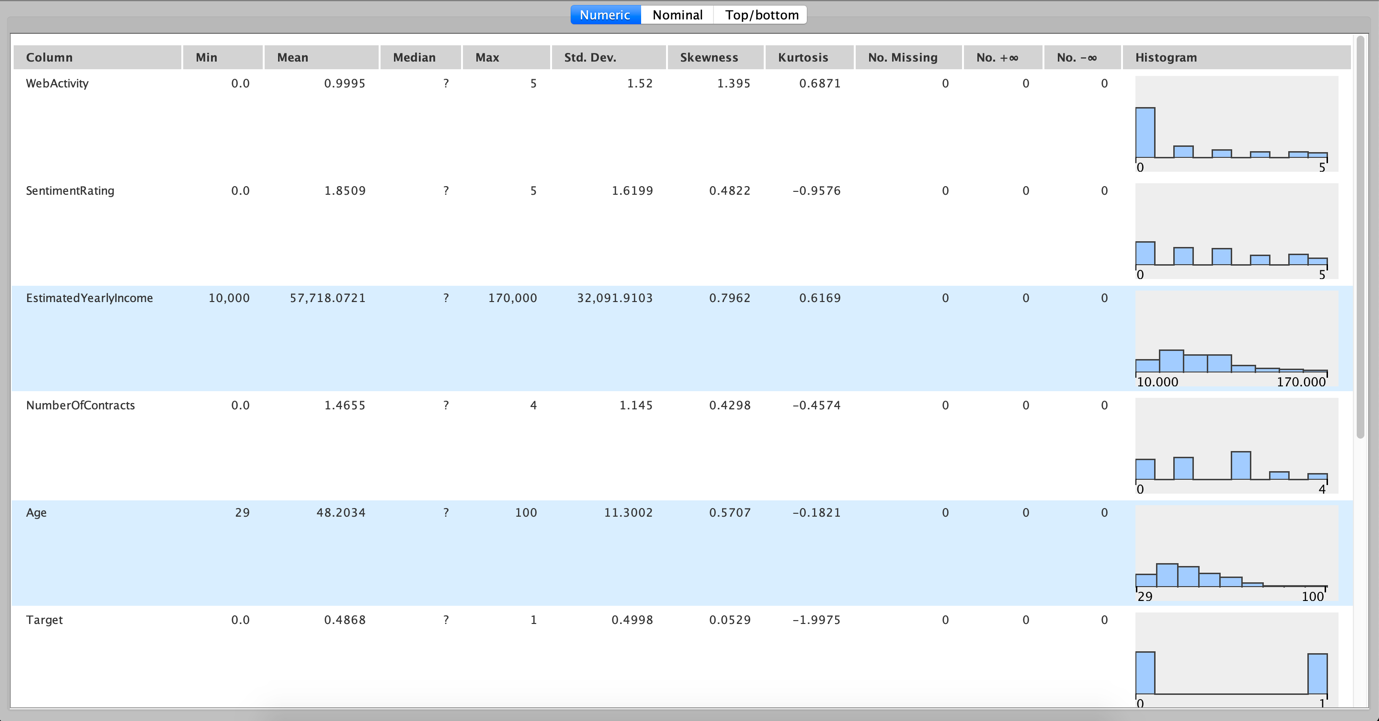


Figure – View do nodo Statistics

A picture containing chart

Description automatically generated

Figure - Matriz de correlação linear

A picture containing chart

Description automatically generated

Figure - Matriz de correlação de classificação

**T2.** Criar plots para visualização dos dados;

Diagram

Description automatically generated

Figure - Vizualização dos dados

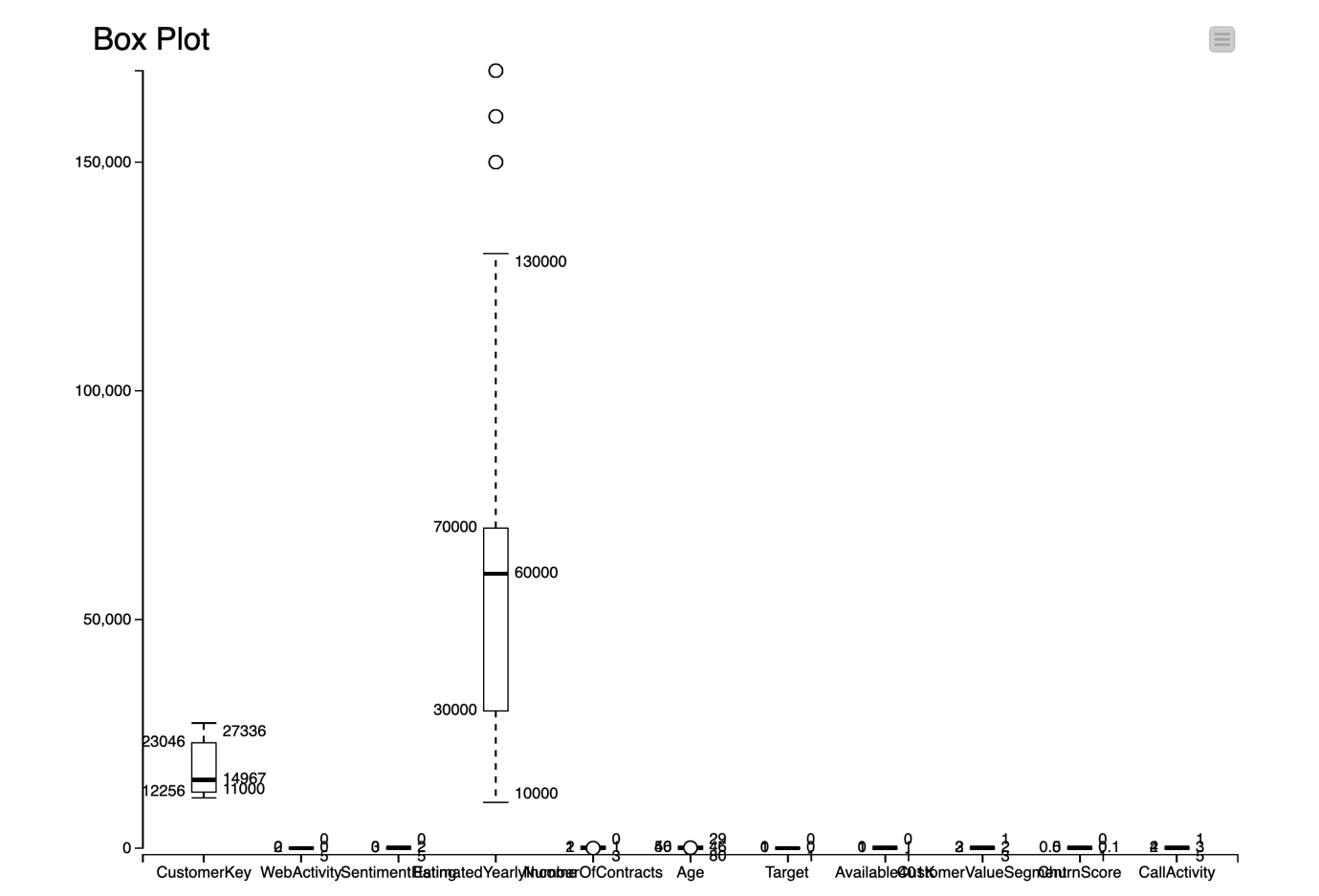


Figure - Box Plot de todos os atributos

Chart, box and whisker chart

Description automatically generated

Figure - Box Plot condicional entre os atributos Age e Gender

**T3.** Aplicar nodos para tratamento de dados de forma a:

A picture containing diagram

Description automatically generated

Figure - Tratamento de dados

a. Excluir todas as colunas do tipo Double;

Graphical user interface, application

Description automatically generated

Figure - Nodo Column Filter

b. Tratar valores em falta;

Graphical user interface

Description automatically generated

Figure - Nodo Missing Value

c. Remover registos duplicados;

Graphical user interface, application

Description automatically generated

Figure - Nodo Duplicate Row Filter

d. Criar 3 bins de igual frequência para a feature age;

Graphical user interface, application

Description automatically generated

Figure - Nodo Auto-Binner

e. Para cada registo, extrair o ano, mês e dia da semana da feature birthday;

Graphical user interface, application

Description automatically generated

Figure - Nodo String to Date&Time

Graphical user interface, application

Description automatically generated

Figure - Nodo Extract Date&Time Fields

f. Excluir utilizadores da plataforma que tenham uma atividade na plataforma

(WebActivity) inferior a 1 hora e que tenham mais de 70 anos;

Graphical user interface, text, application

Description automatically generated

Figure - Nodo Rule-based Row Filter

g. Excluir todos os registos que contenham a sub-string “co” no produto.

Graphical user interface, application

Description automatically generated

Figure - Nodo Row Filter

**T4.** Aplicar nodos para agregação de dados de forma a:

Diagram

Description automatically generated

Figure - Agregação de dados

1. Por género, obter o número e a percentagem de registos, assim como a média da

idade e da atividade na plataforma. Obter também o mínimo e máximo da idade;

Graphical user interface, application

Description automatically generated

Figure - Nodo GroupBy

Graphical user interface, text, application

Description automatically generated

Figure - Nodo GroupBy

1. Por género e atividade na plataforma, obter a moda da análise do sentimento em

relação à plataforma e a média da avaliação do sentimento;

Graphical user interface, application

Description automatically generated

Figure - Nodo GroupBy

Graphical user interface, text, application

Description automatically generated

Figure - Nodo GroupBy

1. Por análise de sentimento, obter o número de registos, a média do salário anual

estimado, o somatório do salário anual e a média do número de contratos.

Graphical user interface, application

Description automatically generated

Figure - Nodo GroupBy

Graphical user interface, text, application

Description automatically generated

Figure - Nodo GroupBy

**T5.** Análise crítica à informação extraída das agregações efetuadas na tarefa anterior. Que conclusões poderia a empresa tirar?

Analisando a informação extraída a partir das agregações efetuadas, algumas das conclusões são:

* O gênero masculino tem uma maior percentagem de registos, apesar da diferença ser pouca;
* Os gêneros têm médias de idades muito próximas (aproximadamente 48 anos);
* O gênero feminino passa, em média, mais tempo na plataforma, apesar da diferença ser pouca;
* O gênero feminino tem o utilizador com maior idade (100 anos);
* A idade do utilizador mais novo é a mesma para ambos os gêneros (29 anos);
* A média do número de contratos de um utilizador aparenta ser proporcional à média do seu salário anual;
* As análises de sentimento extremas correspondem às médias de salário anual extremas, isto é, a pior análise corresponde à menor média salarial e a melhor análise à maior média salarial.

**T6.** Carregar, no Knime, o dataset descarregado. Explorar os dados, procurar informação relevante e mostrar essa mesma informação. P.e., qual a equipa mais indisciplinada? Qual o top-10 dos assistentes para golo? Qual o top-5 de nacionalidades na liga?

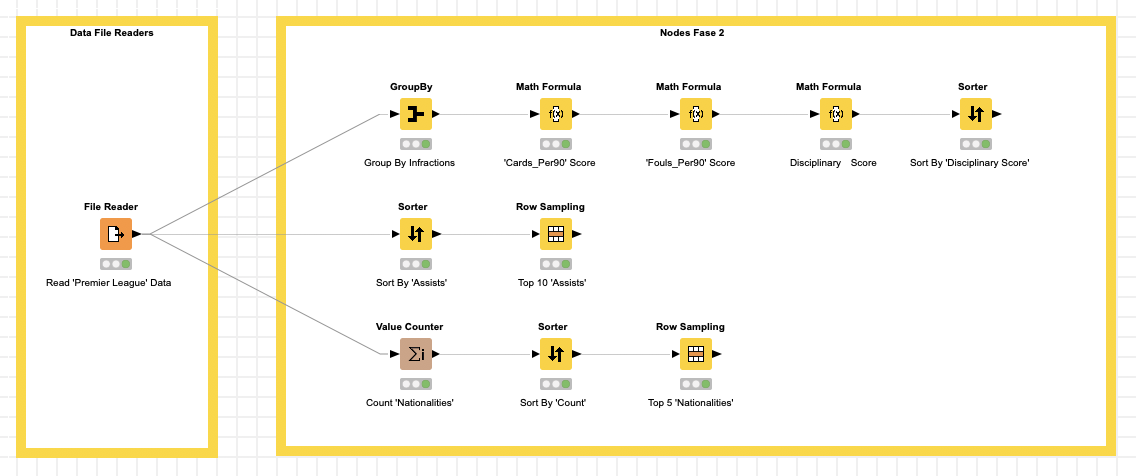


Figure - Exploração dos dados

De modo a obter a equipa mais indisciplinada foi feito um agrupamento por equipas, realizando o somatório das faltas e cartões por 90 minutos. De seguida foram utilizados dois nodos *Math Formula* de modo a multiplicar os somatórios por 0.3 e 0.7 respetivamente de modo a dar maior relevância aos cartões. Por fim os resultados são somados para obter um score e a tabela é ordenada da equipa mais indisciplinada para a menos indisciplinada.

Graphical user interface, application

Description automatically generated

Figura 26 - Nodo GroupBy

Graphical user interface, text, application

Description automatically generated

Figura 17 - Nodo GroupBy

Graphical user interface, application

Description automatically generated

Figura 28 - Nodo Math Formula

Graphical user interface, application

Description automatically generated

Figura 29 - Nodo Math Formula

Graphical user interface, application

Description automatically generated

Figura 30 - Nodo Math Formula

Graphical user interface, application

Description automatically generated

Figura 31 - Nodo Sorter

Table

Description automatically generated

Figura 32 - Tabela final

De modo a obter o top 10 dos assistentes para golo a tabela foi ordenada decrescentemente pelo atributo *assists* e de seguida foi recolhida uma amostra das 10 primeiras linhas.

Graphical user interface, application

Description automatically generated

Figura 3 - Nodo Sorter

Graphical user interface, application

Description automatically generated

Figura 3 - Nodo Row Sampling

Graphical user interface, application, table

Description automatically generated

Figura 3 - Tabela final

De modo a obter o top 5 nacionalidades na liga utilizou-se um nodo Value Counter para contar o número de ocorrências de cada nacionalidade, de seguida a tabela foi ordenada decrescentemente e foi recolhida uma amostra das 5 primeiras linhas.

Graphical user interface, text, application

Description automatically generated

Figura 3 - Nodo Value Counter

Graphical user interface, application

Description automatically generated

Figura 3 - Nodo Sorter

Graphical user interface, text, application

Description automatically generated

Figura 3 - Nodo Row Sampling

Table

Description automatically generated

Figura 3 - Tabela final