Universidade do Estado do Amazonas
Escola Superior de Tecnologia
Data: 2 de outubro de 2020
Disciplina: Fundamentos da Ciência de Dados
Professor: Carlos Maurício S. Figueiredo
Aluno: Marcos Wenneton Vieira de Araújo

Atividade Avaliativa 4
Estudo de Caso no Orange

Introdução

Orange é um kit de ferramentas de visualização de dados de código aberto, Aprendizado de Máquina e mineração de dados. Possui um front-end de programação visual para análise exploratória de dados e visualização interativa de dados. Neste trabalho, serão demonstradas algumas de suas funcionalidades utilizando o conjunto de dados Heart Disease UCI.

Análise Exploratória no Orange

Na Figura 1 encontra-se o esquema criado no orange para a realização da análise exploratória de dados. No escopo deste trabalho serão discutidas as finalidades de alguns dos nós encontrados na imagem.

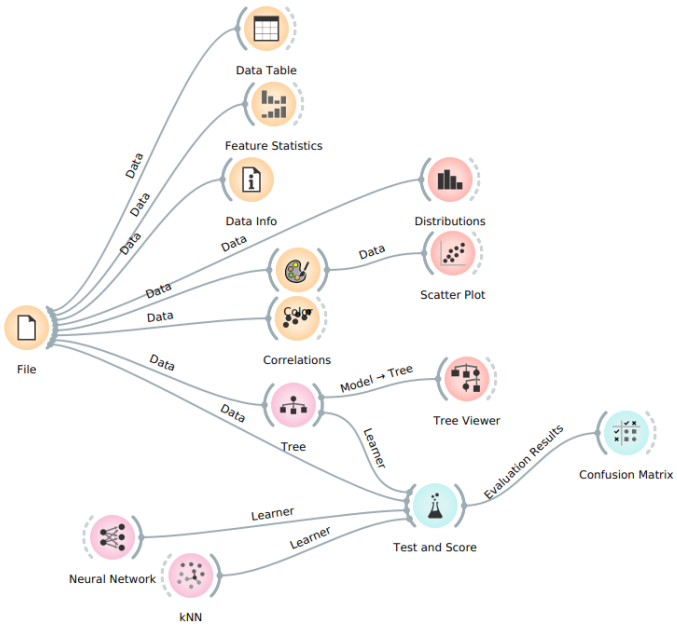

Figura 1: Esquema final da análise exploratória no Orange

Informações do conjunto de dados

O bloco *Data Info* foi utilizado para encontrar informações mais gerais sobre o dataset. Estas informações estão disponíveis abaixo:

- Nome: heart_disease
- Linhas: 303
- Colunas; 14
- Atributos:
  - Categóricos: 7
  - Numéricos: 6
- Atributo Alvo: Categórico com 2 valores

Atributos do Conjunto de Dados

A lista abaixo mostra uma breve descrição sobre cada atributo do dataset, para um melhor entendimento sobre cada um deles:

- **age**: A idade do indivíduo em anos;
- **gender**: O gênero da pessoa (*male* ou *female*);
- **chest pain**: O tipo de dor no peito experimentada (valor 1: *typical angina*, valor 2: *atypical angina*, valor 3: *non-anginal pain*, valor 4: *asymptomatic*);
- **rest SBP**: A pressão sanguínea no indivíduo em mmHg no momento de admissão ao hospital;
- **cholesterol**: O colesterol apresentado pelo indíviduo em mg/dl;
- **fasting blood sugar**: O nível de açúcar no sangue em jejum do indíviduo (¿ 120 mg/dl, 1 = verdadeiro; 0 = falso);
- **rest ECG**: Medida do eletrocardiograma em repouso (valor 1: *normal*, valor 2: *ST-T abnormal*, valor 3: *left vent hypertrophy*);
- **max HR**: A maior medida de frequência cardíaca atingida pelo indivíduo;
- **exerc ind ang**: Angina induzida por exercício (1 = sim; 0 = não)
- **ST by exercise**: depressão no segmento ST (relativo ao eletrocardiograma) induzida por exercício;
- **slope peak exc ST**: a inclinação do segmento ST de pico do exercício (valor 1: *upsloping*, valor 2: *flat*, valor 3: *downsloping*);
- **major vessels colored**: o número de vasos principais (0-3);
- **thal**: Uma anomalia do sangue chamada thalassemia (valor 1: *normal*; valor 2: *fixed defect*; valor 3: *reversable defect*);
- **diameter narrowing (atributo alvo)**: Indica se o indivíduo tem doença no coração (0 = não, 1 = sim).

Estatísticas dos dados

Na Figura 2 podemos ver algumas estatísticas sobre os atributos do conjunto de dados. Nos histogramas apresentados, podemos observar que a cor azul corresponde a pacientes que não possuem doença no coração, enquanto que a cor vermelha indica o contrário.

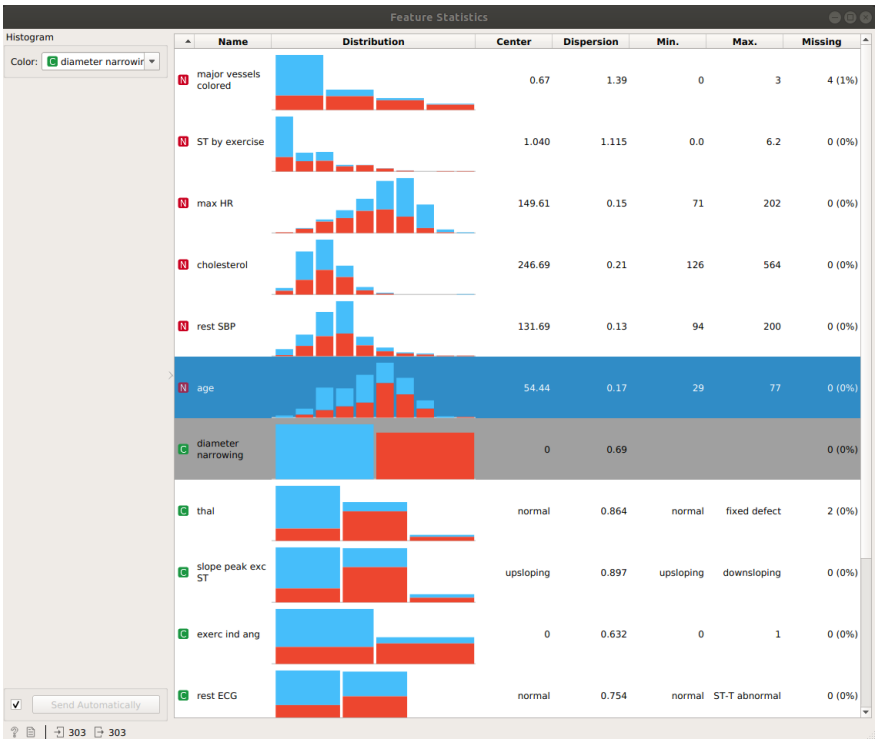

Figura 2: Estatísticas dos atributos do conjunto de dados

Observa-se também que apenas os atributos *major vessels colored* e *thal* possuem informações faltantes, apenas um pouco menos do que 1% dos dados não possuem esses atributos.

Correlação entre atributos

Foi calculada ainda a correlação entre os atributos numéricos disponíveis no dataset. A Figura 3 mostra estas informações, que está disponível no bloco *Correlations* do esquema criado no Orange. Para este bloco, é possível ainda escolher entre o cálculo da correlação de Pearson ou de Spearman.

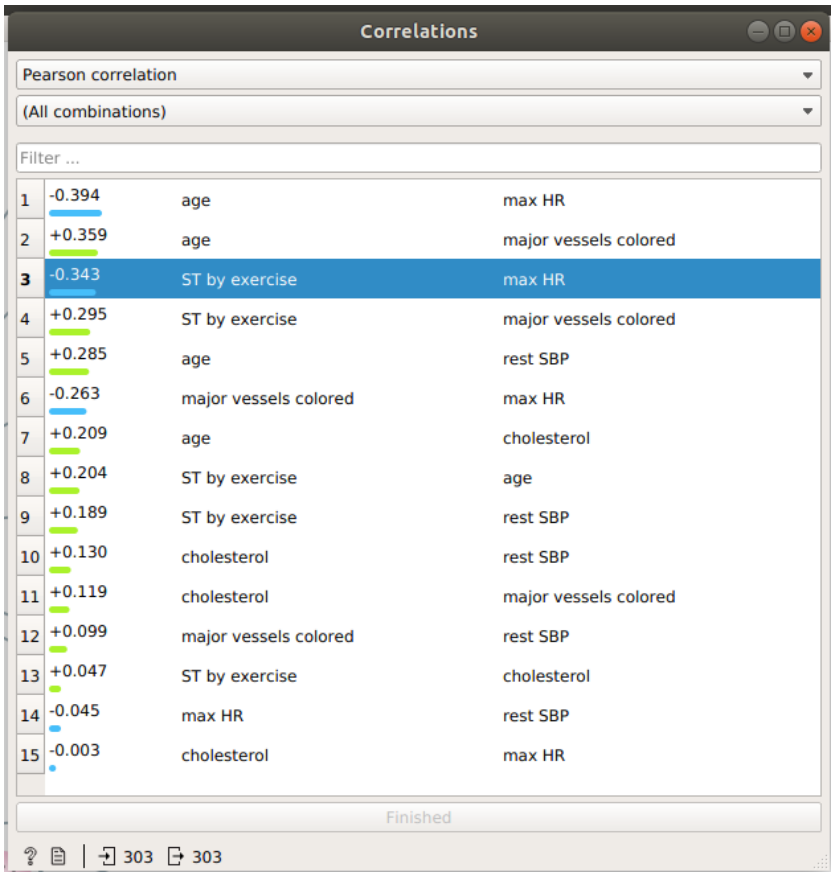

Figura 3: Correlação entre atributos calculadas pelo Orange

Visualização dos Dados

Algumas opções de visualização foram selecionadas, estas estão disponíveis nos blocos *Distributions* e *Scatter Plot*.

Na Figura 4 podemos visualizar um gráfico de distribuição de dados, que permite a escolha do atributo ao qual se deseja obter mais detalhes. Neste caso, as cores da imagem correspondem ao atributo alvo e as quantidades mostradas dizem respeito ao atributo *age*. Com este gráfico, é fácil visualizar que a quantidade de indivíduos doentes entre 55 e 65 anos neste dataset é maior que a quantidade de indivíduos saudáveis.

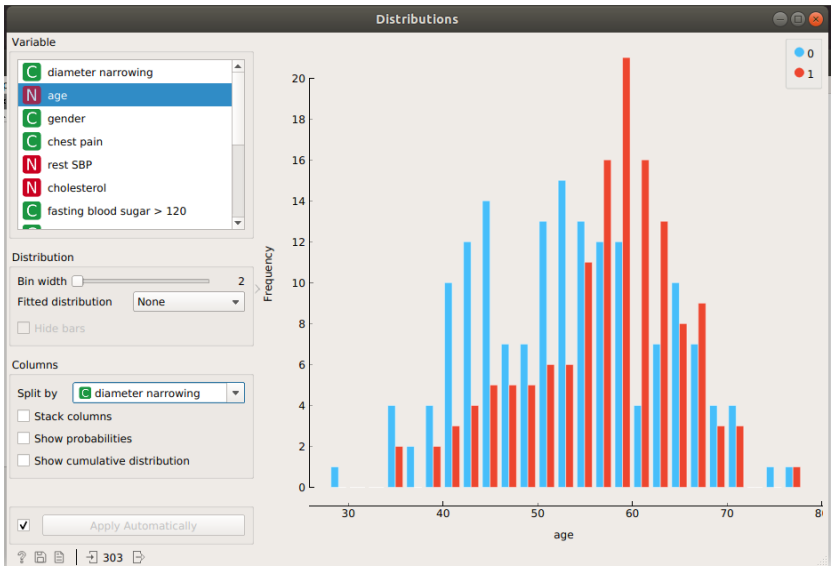

Figura 4: Quantidade de dados do atributo *age*

A Figura 5 por sua vez, mostra um gráfico de dispersão considerando dois atributos numéricos distintos.

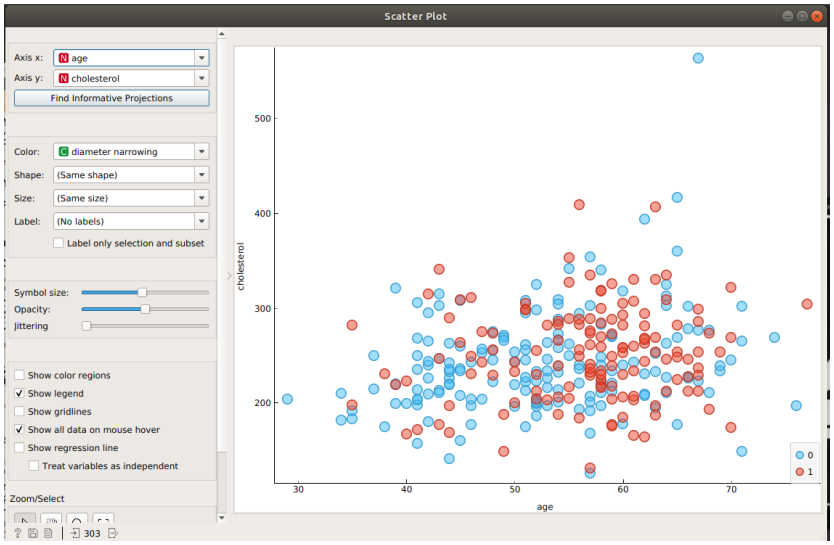

Figura 5: Gráfico de dispersão dos dados considerando os atributos *age* e *cholesterol*

Modelos e Resultados

Ainda foi possível a elaboração de alguns modelos de Aprendizado de Máquina com este dataset. Três modelos distintos foram criados, considerando as técnicas de Árvores de Decisão, Redes Neurais e kNNs.

A Figura 6 mostra uma captura de tela do bloco *Test and Score* na qual estão dispostos os resultados destes modelos. É possível verificar que, dentre eles, o que soube criar uma previsão melhor dos dados foi a Rede Neural, com um *F1-score* de 0.799.

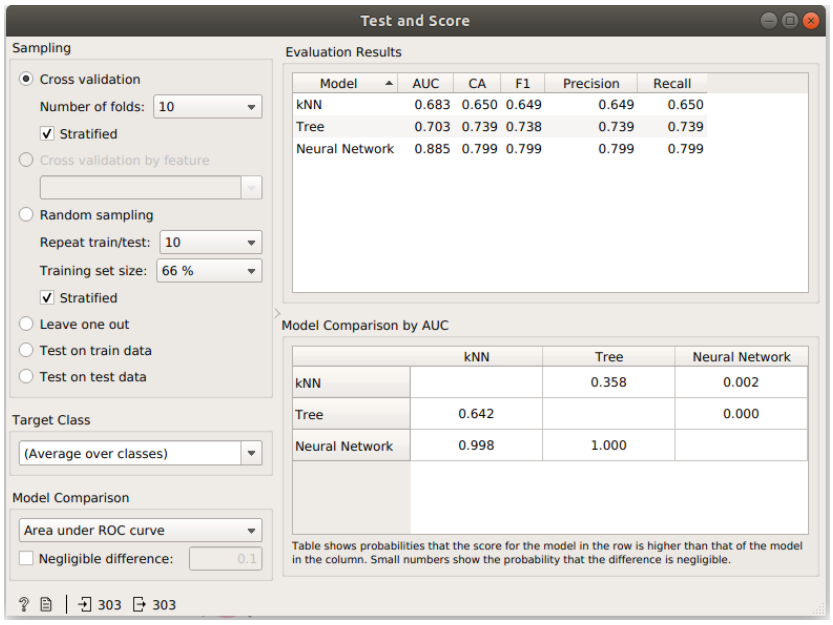

Figura 6: Resultados dos modelos de Aprendizado de Máquina

Para uma melhor verificação desses resultados, foram geradas também matrizes de confusão destes modelos, mostrada na Figura 7. Nelas podemos visualizar, por exemplo, informações de falsos positivos e falsos negativos que foram previstos pelos modelos.

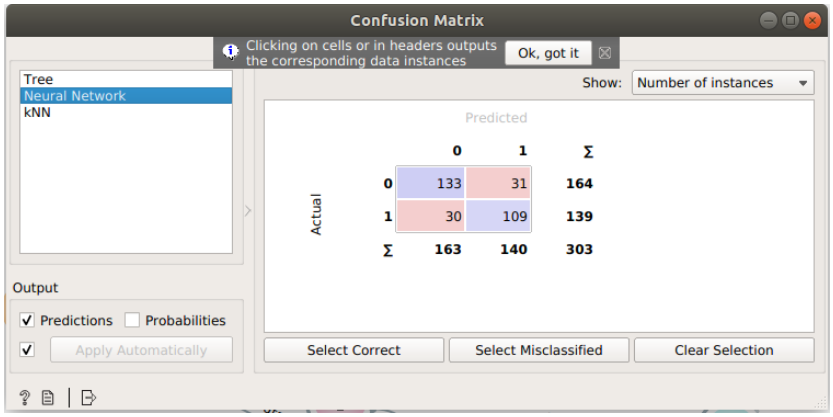

Figura 7: Matriz de Confusão do modelo de Rede Neural