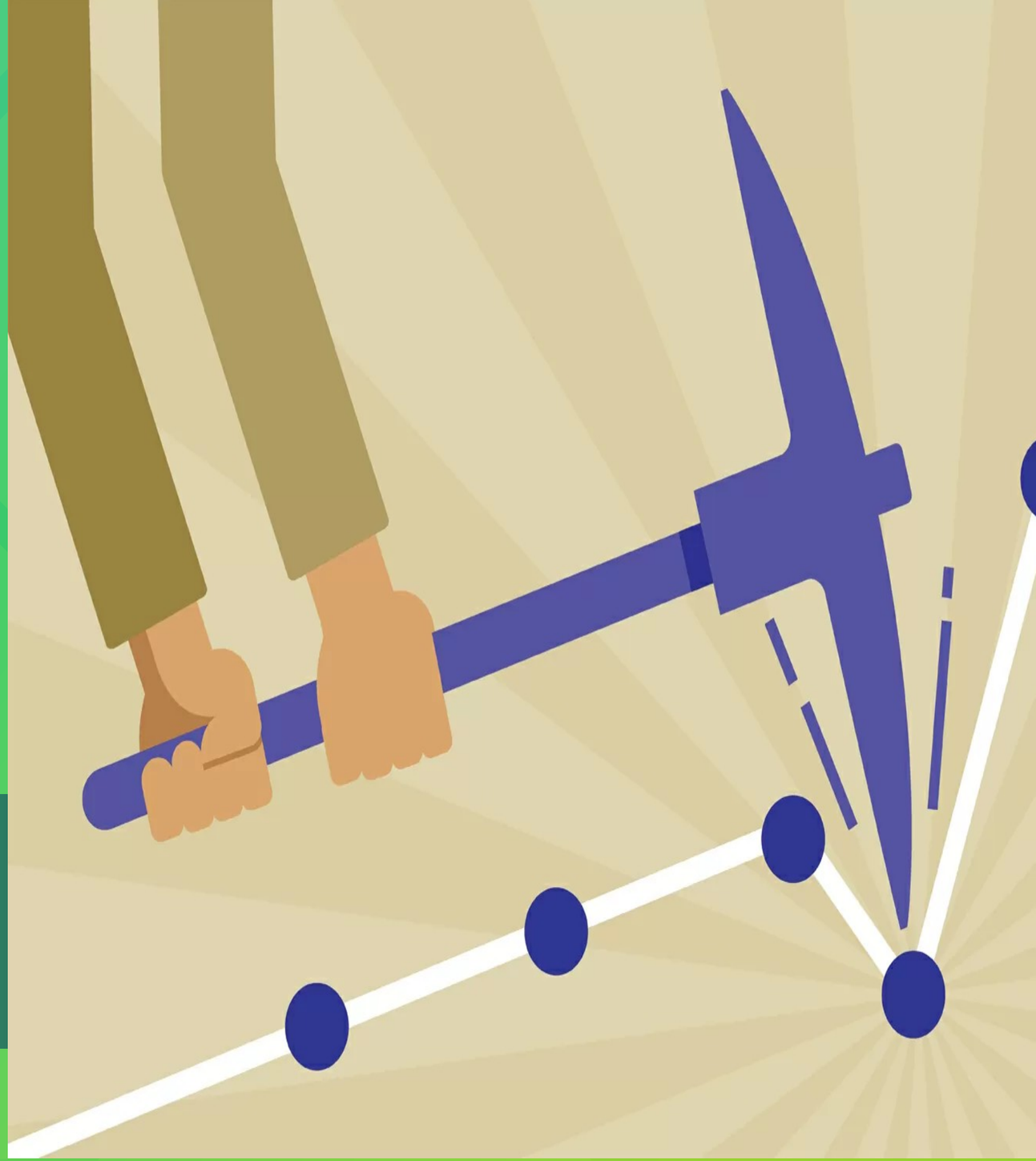
CÂMPUS

Bagé

TÓPICOS EM ADS II



INSTITUTO FEDERAL
Sul-rio-grandense

muito a descobrir

RODRIGO R SILVA

WEKA - Software de Mineração de Dados

- Waikato Environment for Knowledge Analysis;
- Desenvolvido pela Universidade de Waikato (Nova Zelândia);
- Ferramenta livre e open source.

O que é o WEKA?

- Software de mineração de dados e aprendizado de máquina
- Interface gráfica intuitiva (GUI)
- Também permite uso via linha de comando e Java
- Utiliza principalmente o formato ARFF, mas também suporta CSV

WEKA - Software de Mineração de Dados

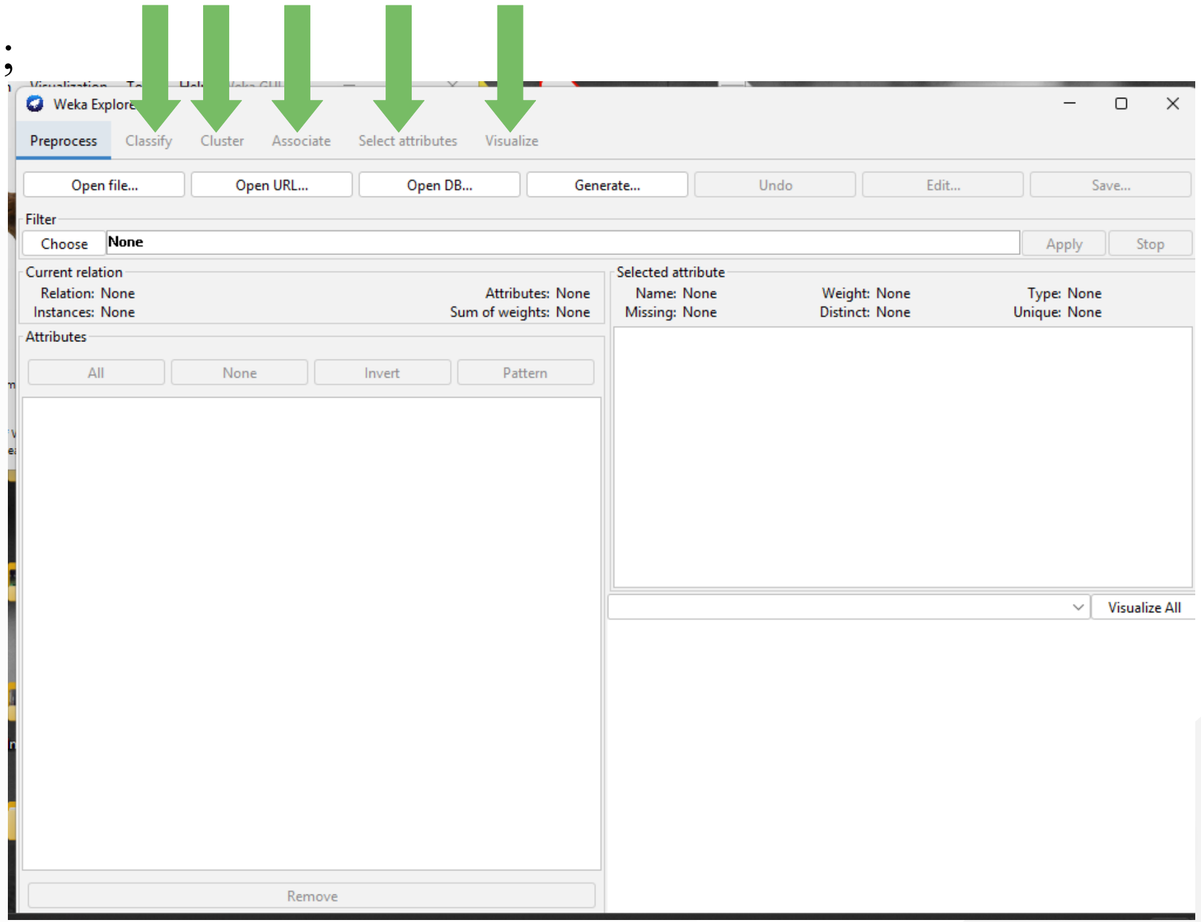
Principais Características

- Conjunto amplo de algoritmos de ML (classificação, regressão, clustering, associação);
- Pré-processamento e seleção de atributos;
- Suporte a avaliação de desempenho (acurácia, F1-score, matriz de confusão);
- Visualização gráfica de dados e resultados.

WEKA - Software de Mineração de Dados

Componentes do WEKA

- Preprocess → preparação e transformação de dados;
- Classify → algoritmos de classificação e regressão;
- Cluster → agrupamento (ex.: K-Means, EM);
- Associate → regras de associação (ex.: Apriori);
- Select Attributes → seleção de variáveis;
- Visualize → visualização gráfica dos resultados;



WEKA - Software de Mineração de Dados

Vantagens do WEKA

- Gratuito e de fácil utilização;
- Grande variedade de algoritmos já implementados;
- Comunidade ativa e documentação extensa;
- Integração com Java e bibliotecas externas.

Limitações do WEKA

- Melhor desempenho em conjuntos de dados pequenos/médios;
- Não é otimizado para Big Data;
- Interface gráfica pode ser limitada em análises muito complexas;
- Requer preparo prévio dos dados.

WEKA - Software de Mineração de Dados

Vantagens do WEKA

- Gratuito e de fácil utilização;
- Grande variedade de algoritmos já implementados;
- Comunidade ativa e documentação extensa;
- Integração com Java e bibliotecas externas.

Limitações do WEKA

- Melhor desempenho em conjuntos de dados pequenos/médios;
- Não é otimizado para Big Data;
- Interface gráfica pode ser limitada em análises muito complexas;
- Requer preparo prévio dos dados.

WEKA – Arquivo ARFF

@RELATION nome_da_base

@ATTRIBUTE atributo1 NUMERIC

@ATTRIBUTE atributo2 {sim,nao}

@ATTRIBUTE atributo3 STRING

@DATA

5.4,sim,"texto exemplo"

6.7,nao,"outra instância"

WEKA – Arquivo ARFF

Um arquivo ARFF segue uma sintaxe padronizada:

```
arff

@RELATION nome_da_base

@ATTRIBUTE atributo1 NUMERIC
@ATTRIBUTE atributo2 {sim,nao}
@ATTRIBUTE atributo3 STRING

@DATA
5.4,sim,"texto exemplo"
6.7,nao,"outra instância"
```

WEKA – ARFF/ Componentes em Detalhe

@RELATION

Define o nome do conjunto de dados.

Exemplo:

```
@RELATION clima
```

WEKA – ARFF/ Componentes em Detalhe

@ATTRIBUTE

Define cada atributo do dataset, incluindo nome e tipo de dado.

Tipos suportados:

- NUMERIC → valores contínuos (ex.: altura, peso).
- INTEGER → valores inteiros.
- REAL → valores reais (aceita decimais).
- STRING → textos livres.
- NOMINAL → valores categóricos enumerados, entre chaves {}.
- Exemplo: {sol,chuva,nublado}
- Exemplo de definição de atributos:

```
@ATTRIBUTE temperatura NUMERIC
@ATTRIBUTE umidade {alta,baixa}
@ATTRIBUTE vento {sim,nao}
@ATTRIBUTE jogar {sim,nao}
```

WEKA – ARFF/ Componentes em Detalhe

@DATA

Marca o início da seção com os valores das instâncias.

Cada linha representa um registro, seguindo a ordem definida em @ATTRIBUTE.

Exemplo:

```
@DATA
30,alta,sim,nao
22,baixa,nao,sim
27,alta,sim,sim
```

WEKA – ARFF / Exemplo

```arff
@RELATION clima

@ATTRIBUTE temperatura NUMERIC
@ATTRIBUTE umidade {alta,baixa}
@ATTRIBUTE vento {sim,nao}
@ATTRIBUTE jogar {sim,nao}

@DATA
30,alta,sim,nao
22,baixa,nao,sim
27,alta,sim,sim
25,baixa,nao,sim
```

WEKA – ARFF

Vantagens do ARFF

- Simplicidade de edição (pode ser aberto em qualquer editor de texto).
- Suporte direto a atributos nominais (categorias).
- Integração nativa com os algoritmos do WEKA.
- Legibilidade humana e compatibilidade com CSV.

Limitações

- Não é otimizado para grandes volumes de dados (Big Data).
- Necessidade de padronizar manualmente os valores (cuidado com maiúsculas/minúsculas e categorias diferentes).
- Não aceita tipos de dados mais complexos diretamente (como imagens ou áudio).