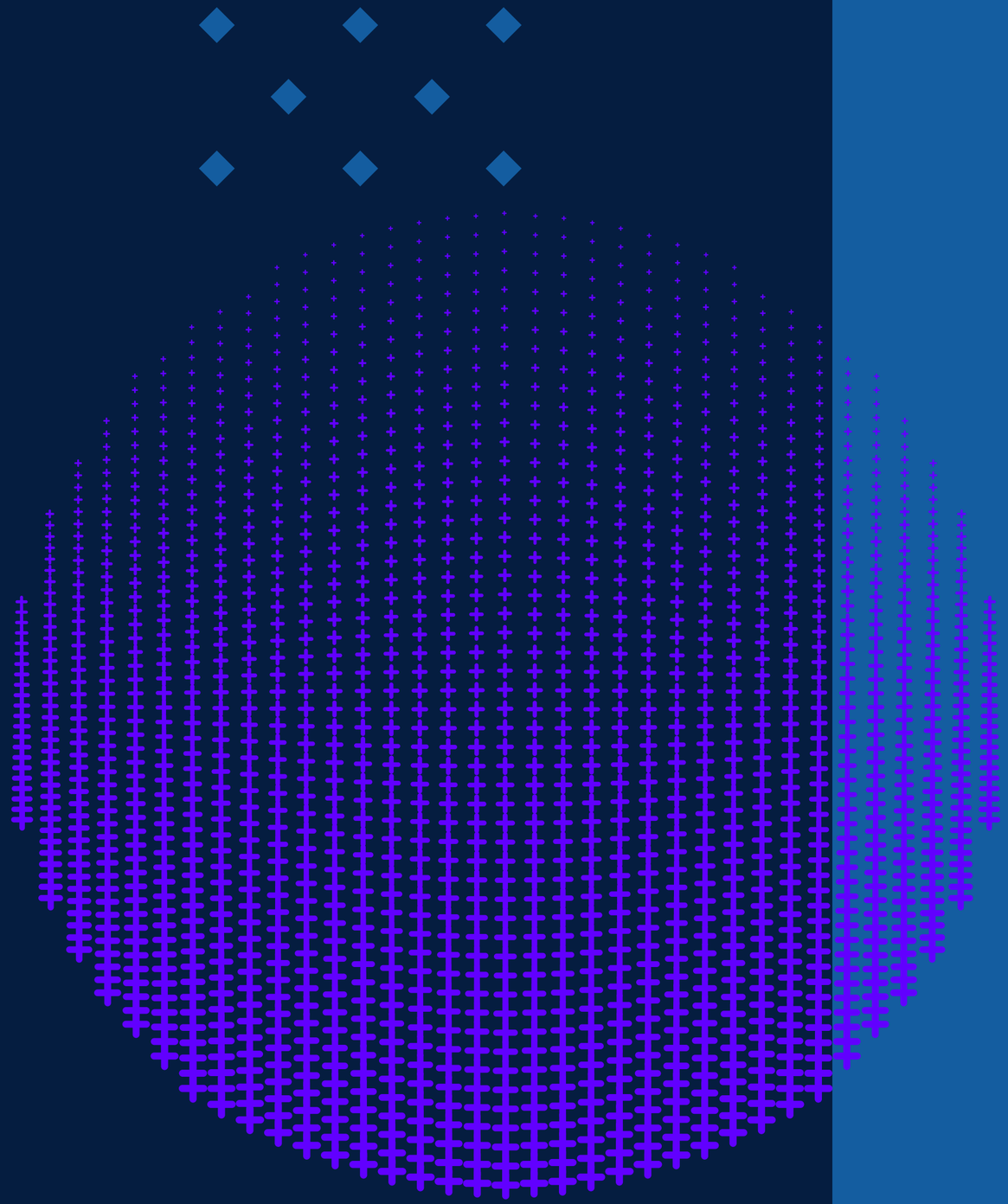
Ciência da Computação 2024.2

EXTRAÇÃO DE CONHECIMENTO E MINERAÇÃO DE DADOS

Unidade 2 – Soluções de machine learning no Knime e Python

Docente: *Ricardo Roberto de Lima*

Discentes:
Beatriz Almeida de Souza Silva
José Nichollas Leandro



KNIME

MODELO 1: Classificação de peso

- **Dataset:** Obesity Classification Dataset (https://www.kaggle.com/datasets/sujithmandala/obesity-classification-dataset)

- **Algoritmo:** Naive Bayes e SVM

- **Dificuldades:**
  - Overfitting nas tentativas iniciais (Naive Bayes e Random Forest).
  - Erro de ‘missing NominalValue’

- **Causa:** O modelo não conseguia estudar direito os dados por que as colunas ‘Gender’ e ‘Label’ estavam sendo convertidas de string para numérico.

- **Solução:** Remover a conversão de string para numérico.

Nós de classificação do Knime dão erro se só receberem valores numéricos, pois eles precisam de valores nominais.

KNIME

MODELO 2: Classificação de preços de celulares

- **Dataset:** Mobile Price Classification (https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification)
- **Algoritmo:** RandomForest e KNN
- **Dificuldades:**
  - Nomes e organização das colunas
- **Solução:** Uso dos nós ‘Column Renamer’ e ‘Column Resorter’ para organização dos dados.
- **Abordagem e Solução:** Estudo dos dados de hardware de vários dispositivos para prever o preço.

KNIME

MODELO 3: Regressão de idades de caranguejos

- **Dataset:** Crab Age Prediction (https://www.kaggle.com/datasets/sidhus/crab-age-prediction)
- **Algoritmo:** RandomForest
- **Dificuldades:**
  - Escolha do algoritmo ideal para a solução
  - Problema ao encontrar colunas que tivessem mais relações entre si
- **Causa:** Os resultados dos modelos estavam muito a baixo dos esperados, aparentemente as colunas não tinham muitas ligações entre elas
- **Solução:** Utilizar o algoritmo de RandomForest para melhor escolha das features de regressão.

PYTHON

MODELO: Regressão de valor de seguro de saúde

- **Dataset:** Medical Cost Personal Datasets (https://www.kaggle.com/datasets/mirichoi0218/insurance)
- **Algoritmo:** LinearRegression
- **Dificuldades:**
  - Não houveram dificuldades. Apenas foi necessário um pouco de conversão de dados (colunas ‘Sex’, ‘Smoker’ e ‘Region’, pois estavam como object).
- **Plataformas de deploy:** Ngrok e Render.
- **Abordagem e Solução:** Estudo dos dados físicos e sociais do usuário para prever o preço de seu seguro de saúde.
- **Possíveis melhoras observadas:** O XGBoost teria sido mais apropriado para o dataset em questão (as métricas com LinearRegression ficaram medianas).

Video-demonstração: