Programa de Disciplina
Especialização em Ciência dos Dados

Módulo: III

Disciplina: Árvores de Classificação e Regressão

Carga Horária: 15 horas (9h teóricas, 4h práticas e 2h de avaliação)

Ofertante: Departamento de Engenharia de Produção – DEENP/UFOP

Objetivo:
Capacitar o estudante no desenvolvimento e implementação de algoritmos não paramétricos para aprendizado supervisionado baseado em árvores de decisão.

Ementa:
Árvores de decisão. Árvores uni e multivariadas. Árvores de classificação. Árvores de regressão. Técnicas de poda. Extração de regras. Aprendizado de regras. *Random Forests*. Implementação em Python. Implementação em R. Aplicações na indústria siderúrgica.

Conteúdo Programático:

1. Árvores de decisão:
    a) Estrutura da Árvore de decisão.
    b) Discriminação.
2. Dimensão de variáveis dependentes:
    a) Árvores univariadas.
    b) Árvores multivariadas.
3. Árvores de classificação:
    a) Qualidade de ajuste.
    b) Medida de pureza e função de entropia.
    c) Exemplos.
4. Árvores de regressão:
    a) Qualidade de ajuste.
    b) Exemplos.

5. Algoritmos:
    a) CART.
    b) ID3.
    c) C4.5.
6. Agregação *bootstrap*:
    a) *Radom Forests*.
    b) *Overfiting*.
7. Aplicações na indústria siderúrgica:
    a) Caso I - Predição de acidentes ocupacionais.
    b) Caso II - Classificação de amostras de aço via espectroscopia de degradação induzida por laser.
8. Implementação em Python.
9. Implementação em R.

Metodologia:
A parte teórica desta disciplina terá a finalidade do embasamento matemático formal e o posicionamento das ferramentas dentro do conjunto de técnicas de aprendizado de máquina. A parte prática buscará capacitar o aluno para utilizar as bibliotecas existentes e ser capaz de analisar dados através dos métodos CART e *Random Forest*. Serão apresentados também estudo de casos reais sobre a aplicação destas técnicas em problemas da indústria siderúrgica.

Atividade Prática Proposta:
Os alunos, em grupo, deverão analisar os dados simulados relativos a um processo de laminação e, a partir das ferramentas estudadas, construir um modelo de classificação de itens defeituosos.

Softwares:

1. Os métodos estudados serão implementados utilizando o Python e o software R.

Bibliografia:

ALPAYDIN, E. *Introduction to Machine Learning*. 3. ed. EUA: MIT Press (MA), 2014. 613 p. ISBN 978-026202-818-9.

IZENMAN, A. J. *Modern Multivariate Statistical Techniques*: Regression, Classification, and Manifold Learning. 2. ed. New York: Springer, 2013. 731 p. ISBN 978-038778-188-4.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002. ISSN 1609-3631.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, n. 85, p. 2825–2830, out. 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>. Acesso em: 15 abr. 2020.

RASCHKA, S.; MIRJALILI, V. *Python machine learning*: Machine Learning and Deeping Learning with Python, scikit-learn, and TensorFlow. 2. ed. Birmingham: Packt Publishing, 2017. 622 p. ISBN 978-178712-593-3.

SARKAR, S. et al. Prediction of occupational accidents using decision tree approach. In: *2016 IEEE Annual India Conference (INDICON)*. Bangalore, India: IEEE, 2016. p. 1–6. ISSN 2325-9418.

ZHANG, T. et al. Classification of steel samples by laser-induced breakdown spectroscopy and random forest. *Chemometrics and Intelligent Laboratory Systems*, v. 157, p. 196–201, 2016. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169743916301630>. Acesso em: 16 abr. 2020.