Bootcamp: Arquiteto(a) de Big Data

Trabalho Prático

Módulo 1: Fundamentos de Big Data

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Coleta de dados.
2. Análise e tratamento de dados.
3. Criar visualização de dados.
4. Implementar algoritmo de Machine Learning.
5. Analisar resultados obtidos.
7. Conhecimento teórico ministrado nas videoaulas.

Enunciado

Um professor do ensino superior propôs um experimento que consistia em desenvolver um modelo preditivo que utilizaria as horas de dedicação aos estudos dos alunos para tentar prever a nota obtida na avaliação final da disciplina.

Para isso, o professor coletou dados de horas dedicadas aos estudos e o resultado final das provas de 100 alunos dos semestres anteriores. Diante disso, o professor observou que a melhor abordagem para esse problema é criar um algoritmo de *Machine Learning* de regressão linear para resolver esse problema.

Portanto, o modelo deve aprender com os dados históricos já coletados e prever o possível resultado obtido a partir das horas de estudo de um novo aluno.

Atividades

Para essa atividade, os alunos deverão criar um algoritmo de regressão linear para prever a nota obtida na avaliação final do aluno baseado no número de horas dedicadas aos estudos.

1. Criar um projeto no Google Drive;
2. Coletar e inserir o arquivo horas_estudo.csv na plataforma;
3. Analisar os dados coletados;
4. Avaliar a relação entre as variáveis;
5. Criar algoritmo de regressão linear;
6. Responder as questões teóricas e práticas do trabalho.

Dicas do professor

1. Analisem com cuidado os dados através da representação gráfica;
2. Analisem bem o gráfico gerado e a disponibilização dos dados;
3. Antes de enviar as respostas, verifiquem se o gabarito está correto;
4. Tenham atenção no que pede cada questão;
5. Os dados disponibilizados no *dataset* são fictícios. Ou seja, não têm relação com o mundo real.
6. Analise os dados do *dataset* e, se for necessário, utilize a função decimal=',' na leitura do arquivo caso o separador dos dados numéricos for um vírgula e não ponto;
7. O *dataset* utilizado no trabalho pode ser obtido no link:

https://github.com/ProfLeandroLessa/classroom-datasets/tree/master/FDA/TP

Bom trabalho prático a todos!