Ciência da Computação / Sistemas de Informação

Avaliação de Algoritmos de Aprendizagem Supervisionada em Dados de Batata Doce para Índices de Produção Considerando o Consumo Humano

Mardem Arantes de Castro - 5º período de computação, UFLA, iniciação científica voluntária

Renato Ramos da Silva - Orientador DCC, UFLA - Orientador(a)

Resumo

A batata-doce é uma importante raiz tuberosa produzida em todo o mundo, é utilizada para consumo humano, animal, produção de combustível e outros fins. Neste estudo, a partir de um experimento conduzido na Universidade Federal de Lavras, no qual houve o plantio de batata-doce, gerou-se um conjunto de dados que seria utilizado junto a técnicas e algoritmos de Aprendizado de Máquina com o objetivo de identificar a combinação ótima de características de uma específica planta que maximiza sua produtividade. A partir do conjunto de dados, inicialmente, foi necessário a limpeza de dados, visto que apresentava uma grande quantidade de variáveis má distribuídas, com grandes quantidades de zeros, distribuições similares entre as outras e também grande quantidade de valores atípicos. Além disso, medidas de pré-processamento foram feitas, com o objetivo de tornar os dados mais ?agradáveis? ao modelo, entre elas, o uso de uma escala logarítmica para reduzir a assimetria de distribuições, codificação para valores categóricos, substituição de valores nulos. Todas as decisões tomadas nessas etapas foram feitas conforme uma análise exploratória de dados. Com o conjunto de dados preparado para o modelo, utilizou-se três algoritmos já consolidados na literatura: Random Forest, Support Vector Machine(SVM) e Lasso Regression. A escolha foi feita com base em uma análise prévia dos dados utilizando diversos outros algoritmos, e esses foram os que se destacaram. Com isso, uma seleção de melhores hiperparâmetros para os modelos foi feita, de forma aleatória. E com os melhores hiperparâmetros, o algoritmo SVM performou melhor. Por fim, a seleção dos melhores atributos foi feita para minimizar o tempo de treinamento e melhorar o resultado do modelo. Como resultados, observou-se a raiz quadrada do erro médio como 4656,85. Além disso, a performance ótima foi observada com 71 atributos (o conjunto de dados possuía 144 atributos). Outro resultado importante foi a obtenção de resultados próximos utilizando apenas 20 atributos, com o custo de apenas 8% do melhor resultado. Esse projeto apresenta uma contribuição de algoritmos de aprendizado de máquina para exploração de dados pós-colheita de batata doce para consumo humano. Como êxito principal desse trabalho, observa-se predições corretas com um número relativamente pequeno de atributos. Ademais, o trabalho foi aceito para a publicação na revista RITA.

Palavras-Chave: Machine Learning, Batata-doce, Aprendizado Supervisionado.
Link do pitch: https://youtu.be/azdFR6xPJzE