Instituto Politécnico
de Castelo Branco
Escola Superior
de Tecnologia

Aprendizagem Computacional no apoio à Gestão Agrícola
Projeto 1

Licenciatura em Informática e Multimédia

Miguel Alexandre Salvado Magueijo
N.º 20191374

Orientadores
Professora Doutora Ana Paula Neves Ferreira da Silva
Professor Doutor Arlindo Ferreira da Silva

Fevereiro 2024

Instituto Politécnico
de Castelo Branco
Escola Superior
de Tecnologia

Aprendizagem Computacional no apoio à Gestão Agrícola
Projeto 1

Miguel Alexandre Salvado Magueijo

N.º 20191374

Orientadores

Professora Doutora Ana Paula Neves Ferreira da Silva

Professor Doutor Arlindo Ferreira da Silva

Trabalho de Projeto apresentado à Escola Superior de Tecnologia do Instituto Politécnico de Castelo Branco para cumprimento dos requisitos necessários à obtenção do grau de Licenciado em Informática e Multimédia, realizada sob a orientação científica do Professor Adjunto Doutor Ana Paula Neves Ferreira da Silva e coorientação do Professor Adjunto Doutor Arlindo Ferreira da Silva, do Instituto Politécnico de Castelo Branco.

Fevereiro 2024

Composição do júri

Presidente do júri

Doutor, Carlos Manuel de Oliveira Alves

Professor Adjunto da Escola Superior de Tecnologia de Castelo Branco

Orientador

Doutor, Ana Paula Neves Ferreira da Silva

Professor Adjunto da Escola Superior de Tecnologia de Castelo Branco

Arguente

Doutor, José Carlos Meireles Monteiro Metrôlho

Professor Coordenador da Escola Superior de Tecnologia de Castelo Branco

Agradecimentos

Num primeiro agradecimento, expresso a minha profunda gratidão à minha família por todo o apoio ao longo dos quase cinco anos de estudo no Instituto Politécnico de Castelo Branco. Um agradecimento especial aos meus pais, Carlos Magueijo e Elsa Magueijo, por acreditarem em mim e por me proporcionarem a oportunidade de dar continuidade aos estudos, especialmente, permitindo a elaboração deste projeto.

Em segundo lugar, quero deixar o meu agradecimento aos professores Ana Paula Silva e Arlindo Silva por toda a orientação e ajuda ao longo do desenvolvimento do projeto, nomeadamente às diversas respostas a todas as minhas perguntas e curiosidades. Quero também agradecer por me ajudarem a entrar no mundo da área de Inteligência Artificial e *Machine Learning* através da realização deste trabalho.

Por fim, quero expressar o meu agradecimento a diversos colegas que me apoiaram em todo este longo percurso. Gostaria de destacar três colegas, André Duarte, Clara Aidos e Pedro Mendonça, que contribuíram de diversas formas para uma melhoria continua do trabalho aqui apresentado.

Resumo

O setor agrícola é um setor muito dependente das condições climáticas. A temperatura e o acesso a água influenciam diretamente o desenvolvimento das culturas. Com o agravamento das alterações climáticas, determinadas plantações acabam por não ser produtivas, uma vez que os agricultores tomam decisões com base no histórico de produção. Este trabalho visa apoiar os agricultores nesta tomada de decisão, ao propor a utilização de técnicas de *Machine Learning* (ML) para a criação de modelos inteligentes com o objetivo de conseguirem recomendar culturas com base em determinadas características do solo e/ou clima.

De forma a ser possível utilizar modelos de ML para a recomendação de culturas, foi necessário estudar as áreas de Inteligência Artificial (IA) e ML dedicadas à criação de sistemas capazes de desempenhar tarefas humanas. Assim, foram estudados os diferentes tipos de aprendizagem computacional e identificado aquele que melhor se enquadra nos objetivos do trabalho, a aprendizagem supervisionada.

Para se perceber o trabalho de investigação disponível na literatura dedicado à resolução deste problema foi realizado um estado da arte do uso de técnicas de ML e *Deep Learning* (DL) na recomendação de culturas. Este estudo permitiu constatar um interesse crescente neste tema nos últimos anos. Adicionalmente, observou-se, de forma positiva, a capacidade de os modelos inteligentes recomendarem culturas corretamente.

O trabalho desenvolvido implicou também a procura de *datasets* adequados para serem utilizados no treino dos modelos. Esta procura evidenciou uma escassez de dados e uma existência de redundância dos mesmos. Depois de um pré-processamento e análise dos *datasets* identificados, passou-se à fase de treino e avaliação dos modelos, conforme delineado nos objetivos. A aplicação dos modelos nos respetivos *datasets*, permitiu constatar de forma positiva e que os modelos de ML são capazes de efetuar recomendação de culturas com sucesso.

Assim, a realização deste projeto permitiu identificar os algoritmos de ML que produziram os melhores modelos, com ênfase no desempenho do algoritmo Random Forest. Foi ainda possível criar uma prova de conceito, à custa do desenvolvimento de um *website*, onde os utilizadores têm a possibilidade de interagir diretamente com os modelos.

Palavras-chave

Dataset para recomendação de culturas; *Machine Learning*; Recomendação de culturas agrícolas; Modelos de *Machine Learning*.

Abstract

The agricultural sector is highly dependent on climatic conditions. Temperature and access to water directly influence the development of crops. As climate change worsens, certain crops end up not being as productive, since farmers make decisions based on production history. This work aims to support farmers in this decision-making process by proposing the use of Machine Learning (ML) techniques to create intelligent models to be able to recommend crops based on certain soil and/or climate characteristics.

To be able to use ML models to recommend crops, it was necessary to study the areas of Artificial Intelligence (AI) and ML that are dedicated to creating systems capable of performing human tasks. Thus, the different types of computer learning were studied and the one that best suited the objectives of the work, supervised learning, was identified.

In order to understand the research work available in the literature dedicated to solving this problem, a state-of-the-art study was carried out on the use of ML and Deep Learning (DL) techniques in crop recommendation. This study revealed a growing interest in this subject in recent years. In addition, the ability of intelligent models to correctly recommend crops has been positively observed.

The work carried out also involved looking for suitable datasets to be used to train the models. This search revealed a shortage of data and the existence of redundancies. After pre-processing and analysing the datasets identified, the models were trained and evaluated, as outlined in the objectives. The application of the models to the respective datasets made it possible to positively verify that the ML models are capable of successfully recommending crops.

As a result, this project made it possible to identify the ML algorithms that produced the best models, with an emphasis on the performance of the Random Forest algorithm. It was also possible to create a proof of concept by developing a *website* where users can interact directly with the models.

Keywords

Dataset for crop recommendation; Machine Learning; Agricultural crop recommendation; Machine Learning models.

Índice geral

1. Introdução 1
1.1. Enquadramento 1
1.2. Objetivos 2
1.3. Planeamento do projeto 3
1.4. Estrutura do relatório 3
2. Inteligência artificial 5
2.1. Machine Learning 7
2.1.1 Aprendizagem supervisionada 10
2.1.2. Aprendizagem não supervisionada 13
2.1.3. Aprendizagem por reforço 14
2.2. Deep Learning 15
2.3. Aquisição de dados 15
2.4. Avaliação e métricas para problemas de classificação 16
3. Estudo do Estado da arte 21
3.1. Metodologia e processo 21
3.1.1. Propósito e objetivos 21
3.1.2. Fontes de dados 23
3.1.3. Estratégia de pesquisa 23
3.1.4. Critérios de inclusão 23
3.1.5. Critérios de exclusão 24
3.1.6. Extração de dados e análise 25
3.2. Análise dos artigos 26
3.3. Discussão dos resultados 35
3.4. Principais conclusões 39
4. Tecnologias e ferramentas utilizadas 40
4.1. Python 40
4.2. Jupyter Notebook 40
4.3. Scikit-learn 41
4.4. Pandas 42
4.5. NumPy 43
4.6. Matplotlib 43
4.7. Seaborn 44

4.8. PyCharm 45
4.9. GitHub 45
4.10. Bibliotecas adicionais 46
5. *Datasets* 48
5.1. Pesquisa e aquisição 48
5.2. Análise e pré-processamento 49
5.2.1. Análise de AtharvaIngle_CR 56
5.2.2. Análise de RaulSingh_CR 59
5.2.3. Análise de KaranNisar_CR 62
5.2.4. Análise de Manikanta_CR 66
5.3. Combinação e verificação de redundância 71
6. Treino e avaliação dos modelos de ML 78
6.1. Processo 78
6.1.1. Carregar o *dataset* e realizar o pré-processamento 79
6.1.2. Seleção de atributos 80
6.1.3. Treino e avaliação dos modelos de ML 82
6.1.4. Exportação dos melhores modelos 84
6.2. Resultados 87
6.2.1. AtharvaIngle_CR 88
6.2.2. Manikanta_CR 94
7. Prova de conceito 99
8. Conclusões 106
8.1. Desafios e trabalho futuro 107
Referências 109

Índice de figuras

Figura 1 - As três áreas principais de IA 6
Figura 2 – Fluxograma da abordagem de resolução de um problema sem ML 8
Figura 3 - Fluxograma da abordagem de resolução de um problema com ML 9
Figura 4 - Adaptação da abordagem de um problema com ML 9
Figura 5 –Exemplificação de dados de treino para o problema *SPAM* de mensagens em ML 11
Figura 6 - Exemplo de agrupamentos de dados através do uso da técnica de aprendizagem não supervisionada 13
Figura 7 - Matriz de confusão e posições dos VP, VN, FP e FN 18
Figura 8 - Exemplo real de matriz de confusão para classificação de frutas 18
Figura 9 - Exemplo do cálculo da precisão e *recall* 20
Figura 10 - Gráfico da Scopus com número de publicações que utilizam ML para recomendar culturas 22
Figura 11 - Gráfico da ACM com número de publicações que utilizam ML para recomendar culturas 22
Figura 12 - Fluxograma PRISMA resultante do uso da metodologia 24
Figura 13 - Total de menções para cada algoritmo ML utilizado 37
Figura 14 - Total de vezes que o algoritmo ML produziu o melhor modelo 38
Figura 15 – Exemplo de divisão por células de código e *output* da sua execução num ficheiro Jupyter Notebook 41
Figura 16 - Exemplo de um *DataFrame* da biblioteca pandas após ser carregado um ficheiro CSV 42
Figura 17 - Exemplos de estilos de gráficos que é possível criar usando a biblioteca Matplotlib 43
Figura 18 - Comparação do código necessário para um gráfico simples usando as bibliotecas *matplotlib* e *seaborn* 44
Figura 19 – PyCharm Professional Edition, visualização de um objeto 'DataFrame' em tempo real 45
Figura 20 - Exemplo de alterações de um ficheiro entre a penúltima e última versão no GitHub 46
Figura 21 - Primeiros dois blocos de "analyse_dataset.ipynb" 50
Figura 22 - Código de "clean_*dataset*.py" que limpa e guarda o ficheiro do *dataset* escolhido 51
Figura 23 - Distribuição das instâncias por atributo para AtharvaIngle_CR 56
Figura 24 - Gráficos de dispersão para os atributos de AtharvaIngle_CR 58
Figura 25 - Gráfico de dispersão para '*K*' e '*humidity*' de AtharvaIngle_CR 59
Figura 26 - Distribuição das instâncias por atributo para RaulSingh_CR 60
Figura 27 - Gráficos de dispersão para os atributos de RaulSingh_CR 61
Figura 28 – Comparação do gráfico de dispersão para '*K*' e '*humidity*' entre AtharvaIngle_CR e RaulSingh_CR 62
Figura 29 - Distribuição das instâncias por atributo para KaranNisar_CR 63
Figura 30 - Gráficos de dispersão para os atributos de KaranNisar_CR 65

Figura 31 - Gráfico de dispersão para *'P'* e *'rainfall'* de KaranNisar_CR.....66
Figura 32 – Primeira distribuição das instâncias por atributo para Manikanta_CR 67
Figura 33 - Distribuição das instâncias por atributo para o *dataset* Manikanta_CR após remoção dos *outliers*.....68
Figura 34 - Gráficos de dispersão para os atributos de Manikanta_CR.....70
Figura 35 - Gráfico de dispersão para *'B'* e *'K'* de Manikanta_CR.....71
Figura 36 - Código para combinar os diversos ficheiros '.csv' de *datasets*.....72
Figura 37 - Instância duplicada numa primeira combinação de *datasets*.....73
Figura 38 - Instâncias com o mesmo valor nos atributos *ph* e *rainfall* no *dataset* resultante da combinação.....74
Figura 39 - Restantes instâncias do *dataset* combinado após remoção de instâncias malformadas.....75
Figura 40 - Instâncias originais de AtharvaIngle_CR para a cultura *'chickpea'*.....76
Figura 41 - Duplicação de instâncias da cultura *'chickpea'* com alteração da denominação para *'soyabeans'*.....76
Figura 42 - Processo de treino e avaliação em quatro passos.....78
Figura 43 – Etapas constituintes do processo "Carregar o *dataset* e realizar o pré-processamento".....79
Figura 44 - Etapas constituintes do processo "Seleção de atributos".....81
Figura 45 - Etapas constituintes do processo "Treino e avaliação dos modelos de ML".....82
Figura 46 – Representação gráfica da técnica de treino validação cruzada (*Cross-Validation*).....83
Figura 47 - Etapas constituintes do processo "Exportação dos melhores modelos " 85
Figura 48 - Exemplo de um ficheiro metadados criado na exportação de modelos ML.....86
Figura 49 - Exemplo do diretório de exportação com todos os ficheiros dos modelos de ML, metadados e *dataset*.....86
Figura 50 – Representação gráfica do melhor modelo de ML gerado pelo algoritmo Árvore de Decisão para o *dataset* AhtarvaIngle_CR.....92
Figura 51 - Matrizes de confusão para os melhores modelos de ML do conjunto nº3 de Manikanta_CR.....98
Figura 52 - Interface da página inicial do *website*.....100
Figura 53 - Interface da página de interação com os modelos de ML treinados no dataset AtharvaIngle_CR.....101
Figura 54 - Interface da página de interação com os modelos de ML treinados no dataset Manikanta_CR.....101
Figura 55 - Interface para a seleção dos modelos de ML do respetivos *dataset*..102
Figura 56 - Erro apresentado ao utilizador quando a *interface* deteta um erro nos campos.....103

Índice de fórmulas

Fórmula 1 - Cálculo da exatidão 19
Fórmula 2 – Outra formulação possível para o cálculo da exatidão 19
Fórmula 3 - Cálculo da precisão 19
Fórmula 4 - Cálculo da medida *recall* 19
Fórmula 5 - Cálculo da medida *F1-score* 20

Índice de tabelas

Tabela 1 - Cronograma de tarefas do projeto 3
Tabela 2 – Conjunto de dados a serem extraídos de cada artigo analisado 25
Tabela 3 - Ano publicação e *dataset*(s) extraídos de cada artigo 26
Tabela 4 - Extração dos algoritmos de ML e melhor modelo de cada artigo analisado 27
Tabela 5 - Informação extraída de cada *dataset* escolhido 52
Tabela 6 - Descrição e unidade de medida de cada atributo 54
Tabela 7 - Número de instâncias por cultura de cada *dataset* após limpeza 55
Tabela 8 – Tabela com estatísticas das instâncias de AtharvaIngle_CR 57
Tabela 9 - Tabela com estatísticas das instâncias de RaulSingh_CR 60
Tabela 10 - Tabela com estatísticas das instâncias de KaranNisar_CR 64
Tabela 11 –Sinalização das instâncias com valores anómalos de Manikanta_CR . 67
Tabela 12 - Estatísticas das instâncias do *dataset* Manikanta_CR 69
Tabela 13 - Siglas e respetiva correspondências aos algoritmos de ML utilizados 88
Tabela 14 - Conjunto de atributos identificados por RFCEV e SelectFromModel, incluindo o conjunto com todos, para AtharvaIngle_CR 89
Tabela 15 - Resultados do *dataset* AtharvaIngle_CR para o conjunto de atributos nº1 90
Tabela 16 - Resultados do *dataset* AtharvaIngle_CR para o conjunto de atributos nº2 90
Tabela 17 - Resultados do *dataset* AtharvaIngle_CR para o conjunto de atributos nº3 91
Tabela 18 - Comparação entre resultados obtidos e de estado da arte para o *dataset* AtharvaIngle_CR 93
Tabela 19 - Conjunto de atributos identificados por RFCEV e SelectFromModel, incluindo o conjunto com todos, para Manikanta_CR 95
Tabela 20 - Resultados do *dataset* Manikanta_CR para o conjunto de atributos nº1 95
Tabela 21 - Resultados do dataset Manikanta_CR para o conjunto de atributos nº2 96
Tabela 22 - Resultados do *dataset* Manikanta_CR para o conjunto de atributos nº3 96

Lista de abreviaturas, siglas e acrónimos

ANN (*Artificial Neural Network*)

API (*Application Programming Interface*)

BO (*Bayesian Optimization*)

BiLSTM (*Bidirection Long short-term memory*)

CDT (*C4.5 Decision Tree*)

CRNN (*Cascaded Recurrent Neural Network*)

DL (*Deep Learning*)

DNN (*Deep Neural Network*)

EC (*Eletrical Conductivity*)

EFB (*Exclusive Feature Bundling*)

ELM (*Extreme Leraning Machine*)

EO (*Equilibrium Optimizer*)

FFO (*Fruit Fly Optimization*)

FN (Falsos Negativos)

FP (Falsos Positivos)

GBRT (*Gradient Boosted Regression Tree*)

GDP (*Gross Domestic Product*)

GOSS (*Gradient-based One Side Sampling*)

GRU (*Gated Recurrent Unit*)

HMFO-ML (*Hybrid Moth Flame Optimization with Machine Learning*)

IA (Inteligência Artificial)

IBk (*Intance-based learning with parameter k*)

ICRYP-DFADL (*Intelligent Crop Recommendation and Yield Prediction based on Drafonfly Algorithm based Deep Learning*)

IDE (*Integrated Development Environmnet*)

IG (*Information Gain*)

IoT (*Internet of Things*)

IoTSNA-CR (*Internet of Things-enabled soil nutrient classification and crop recommendation*)

KELM (*Kernal Extreme Learning Machine*)

KNN (*K-Nearest Neighbors,* N Vizinhos mais próximos)

LSTM (*Long short-term memory*)

LoR (*Logistic Regression*, Regressão Logistica)

MDPs (*Markov Decision Process*)

ML (*Machine Learning*)

MLP (*Multi-Layer Perceptron*)

MLR (*Multiple Linear Regression*)

MMML-CRYP (*Multimodal Machine Learning Based Crop Recommendation and Yield Prediction*)

MSVM-DAG-FFO (*Multiclass Support Vector Machine with a directed by acyclic graph and Fruit Fly Optimization*)

ONN (*Other Neural Network*)

PART (*Partial C4.5 Decision Tree*)

PCA (*Principal Component Analysis*)

PNN (*Probabilistic Neural Network*)

PRISMA (*Preferred Reporting Systematic Reviews and Meta-Analyses*)

REP (*Reduced Error Pruning*)

RFOERNN-CRYP (*Red Fox Optimization with Ensemble Recurrent Neural Network for Crop Recommendation and Yield Prediction*)

RNN (*Recurrent Neural Network*)

SVC (*Support Vector Classifier*)

SVM (*Suppor Vector Machine*)

SVR (*Support Vecotr Regresion*)

VN (Verdadeiros Negativos)

VP (Verdadeiros Positivos)