INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Melhorar a Sustentabilidade da Irrigação usando Machine Learning

Francisco José dos Santos Negrier Raimundo

Mestrado em Engenharia de Telecomunicações e Informática

Orientador
Professor Doutor Pedro Joaquim Amaro Sebastião, Professor Auxiliar
ISCTE-IUL

Coorientador
Mestre André Filipe Xavier da Glória, Investigador
ISCTE-IUL

Outubro, 2021

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologias da Informação

Melhorar a Sustentabilidade da Irrigação usando Machine Learning

Francisco José dos Santos Negrier Raimundo

Mestrado em Engenharia de Telecomunicações e Informática

Orientador
Professor Doutor Pedro Joaquim Amaro Sebastião, Professor Auxiliar
ISCTE-IUL

Coorientador
Mestre André Filipe Xavier da Glória, Investigador
ISCTE-IUL

Outubro, 2021

À minha mãe, irmã, avó e namorada por me terem dado condições de estudar e por me terem apoiado sempre.

Agradecimentos

Primeiro quero agradecer a toda a minha família, à minha avó, à minha irmã, em especial à minha mãe por me ter sempre dado condições de estudar estes últimos 5 anos e por me incentivar, mesmo que por vezes indiretamente, a superar todos os obstáculos que tive ao longo destes últimos anos e também ao meu avô que apesar de não estar já presente, inspirou-me e fez com que conseguisse chegar até aqui. Quero agradecer ao Professor Pedro Sebastião pelo apoio e orientação ao longo da dissertação.

Quero agradecer ao Mestre André Glória pelo apoio, pela orientação e por estar sempre disponível a resolver qualquer problema que eu enfrentasse em toda a dissertação. Sem ti nunca teria conseguido, o meu muito obrigado.

O meu muito obrigado à Filipa Gouveia, por todo o apoio que me deu ao longo da dissertação, sendo que foi fundamental para nunca em momento algum eu pensasse em desistir.

Resumo

Hoje em dia é de grande importância pouparmos os poucos recursos que ainda existem no planeta Terra, sendo que a água tem um papel fundamental na nossa sobrevivência. Com o desenvolvimento de novas tecnologias, podemos usar a mesmas a nosso favor de forma a contrariar o consumo e o desperdício da água. Recorrendo à internet of things, inteligência artificial e machine learning podemos desenvolver sistemas inovadores e completos que visam efetuar, por um lado, uma poupança do consumo de água e, por outro a adequar a irrigação certa em tempo real e de forma adaptada às necessidades, no sector da agricultura.

Esta dissertação apresenta uma proposta de solução para o desenvolvimento de um sistema recorrendo a mecanismos de machine learning, capaz de fazer a previsão de dados meteorológicos futuros e, através da análise dos mesmos de indicar se há necessidade de rega, indicando a hora e a duração da rega. Ao longo do sistema foram testados vários algoritmos de machine learning, sendo que o que apresentou melhores resultados foi o algoritmo floresta aleatória. Com o recurso a este algoritmo foi possível gerar uma redução do tempo de rega na ordem dos 72%.

Através deste trabalho foi possível mostrar que a utilização de um sistema que é suportado em machine learning permite reduzir a quantidade de água desperdiçada, quando comparado com sistemas sem a utilização de inteligência artificial.

Palavras-Chave: Aprendizagem Automática, Agricultura Sustentável, Previsão, Análise de Dados.

Abstract

Nowadays, it is very important that we save the few resources that still exist on planet Earth, and water has a fundamental role in our survival. With the development of new technologies, we can use them to our advantage to counteract the consumption and waste of water. Using the internet of things, artificial intelligence and machine learning we can develop innovative and complete systems that aim to carry out, on the one hand, a saving of water consumption and, on the other hand, to adapt the right irrigation in real time and in a way adapted to the needs, in the agriculture sector.

This dissertation presents a proposal for a solution and development for a system using machine learning mechanisms, capable of forecasting future meteorological data and, through their analysis, indicate whether there is a need for irrigation, indicating the time and duration of the irrigation. Throughout the system, several machine learning algorithms were tested, and the one with the best results was the random forest algorithm. With the use of this algorithm it was possible to generate a reduction of the watering time in the order of 72%.

Through this work it was possible to show that the use of a system that is supported in machine learning we can reduce the amount of wasted water compared to a conventional method without the use of any intelligence mechanism.

Keywords: Machine Learning, Sustainable Agriculture, Forecast, Data Analysis.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AD: Árvore de Decisão.
API: Interface de Programação de Aplicações.

FA: Floresta Aleatória.

IA: Inteligência Artificial.
IoT: Internet of Things.
IP: Protocolo de Internet.

LoRaWAN: Long Range Wide Area Network.
LTE: Long-Term Evolution.

ML: Machine Learning.

NWP: Previsão Numérica do Tempo.

RNA: Rede Neuronal Artificial.
RSSF: Rede de Sensores sem Fios.

S3VM: Máquina de Vetores de Suporte Semi-Supervisionada.
SVM: Máquina de Vetores de Suporte.

Wi-Fi: Wireless Fidelity.

XGBoost: Extreme Gradient Boosting.

CAPíTULO 1

Introdução

1.1. Motivação

Hoje em dia a sustentabilidade é ligada diretamente à inovação, uma vez que as pessoas estão cada vez mais conscientes das alterações climáticas, escassez da água, energias renováveis e outros desafios, que requerem a existência de novos produtos e abordagens, nomeadamente tecnológicas, que levam a avanços ecológicos. A Internet of Things (IoT) e o Machine Learning (ML) são duas das novas tecnologias que estão a ter um grande impacto nas vidas das pessoas ao longo dos últimos anos, com grande parte da sua investigação a ser feita nas áreas ligadas à sustentabilidade. Através da recolha de dados e da análise dos mesmos em tempo real é possível oferecer ao utilizador uma vantagem face a potenciais perigos ou gastos desnecessários que com uma abordagem mais tradicional não seriam possíveis de atingir. Estas tecnologias no âmbito das cidades inteligentes têm uma grande importância e cada vez se torna mais imperativo serem utilizadas para melhorar toda a sustentabilidade das mesmas. Pode-se conseguir beneficiar destas duas tecnologias principalmente na agricultura, onde existem ainda muitas plantações sem qualquer mecanismo de inteligência ou automação, sendo, em alguns casos, feito de forma convencional. Não existe ainda um acompanhamento ao pormenor das necessidades das plantações, do solo e da irrigação, e isso tem de ser alterado em todo o planeta, uma vez que para os humanos, fazer certas decisões, pode ser uma tarefa com desafio e com alguma complexidade. Mesmo nos casos onde já existe alguma automação, esta não é suficiente, e este trabalho vai tentar impor essa automação necessária em toda a agricultura, sejam hortas de pequenas dimensões ou mesmo grandes extensões de plantações. Neste sentido, surge necessidade de implementar um algoritmo de machine learning para se conseguir fazer uma gestão da água mais sustentável, de forma a que o desperdício de água seja praticamente nulo ou mesmo nulo, uma vez que a água é um recurso fundamental e que cada vez existe menos e é indispensável para a vida de todos nós.

1.2. Objetivos

O objetivo principal deste trabalho é desenvolver um sistema suportado por algoritmo, recorrendo a mecanismos de machine learning e aprendizagem automática, que recebe dados em tempo real de parâmetros como temperatura, humidade do solo, dados meteorológicos da região, consumos, disponibilidade energética, entre outros, de redes de sensores instaladas diretamente em campos agrícolas, ou de estações meteorológicas nas proximidades dos campos agrícolas, de modo a identificar a necessidade de ligar ou desligar o sistema de rega, tendo como finalidade o aumento da eficiência do mesmo, levando a uma redução do consumo de água, consumo de energia e também do custo para o consumidor final. O sistema será desenvolvido de modo a que sempre que cheguem novos dados seja possível ajustar o sistema de rega à melhor eficiência possível. Para ser possível atingir o objetivo principal acima referido, primeiro será tido em conta alguma literatura sobre IoT, machine learning, green technologies e sistemas relacionados que possam existir sobre problemas relacionados. De seguida será feito um estudo de comparação sobre os diferentes tipos de algoritmos de machine learning, sendo que, dos que se mostrarem ter melhores características, se pretende com esses desenvolver as soluções para que se consiga perceber qual ou quais poderão conduzir aos melhores resultados possíveis.

1.3. Contribuições Científicas

Esta dissertação deu origem ao desenvolvimento do seguinte artigo, tendo o mesmo sido aceite e publicado:

- F. Raimundo, A. Glória and P. Sebastião, "Prediction of Weather Forecast for Smart Agriculture supported by Machine Learning,"2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0160-0164.

1.4. Estrutura da Dissertação

A dissertação é dividida por 7 capítulos. Neste primeiro capítulo, foi introduzido o tema e os objetivos desta dissertação. No Capítulo 2 encontra-se o estado da arte, onde se pode perceber definições, tecnologias que existem e perceber que trabalhos foram desenvolvidos na área de interesse deste trabalho e os resultados que os mesmos obtiveram. O Capítulo 3 apresenta a arquitetura do sistema, em que são demonstrados alguns objetivos e como se vai proceder para os alcançar. No Capítulo 4 são testados dois datasets e vários algoritmos de machine learning. De forma a escolher qual o melhor dataset e o algoritmo que melhor consegue prever os dados meteorológicos. Posteriormente são usados

pelo sistema final de modo a calcular os tempos de rega. No Capítulo 5 é demonstrado que através da obtenção e análise de dados se pode calcular o tempo de rega caso, haja necessidade da mesma, utilizando dois tipos de rega e dois métodos para funcionamento do sistema, um método que recorre a machine learning e outro método mais convencional. No Capítulo 6 procedeu-se à validação do sistema, através da comparação de 3 cenários diferentes, de forma a validar qual o que consegue ser mais eficaz e eficiente ao nível da poupança da água e dos custos, tanto ao nível do equipamento como da sua manutenção. E por fim, no Capítulo 7, as conclusões desta dissertação são apresentadas e também o trabalho futuro a desenvolver.

CAPíTULO 2

Revisão da Literatura

Este capítulo tem como base uma revisão da literatura na área de Internet of Things (IoT) e de Machine Learning (ML) com foco na agricultura, de forma a entender as várias abordagens utilizadas em trabalhos anteriores.

Na Secção 2.1 é exposto o conceito de IoT e como este pode ser dividido em vários elementos. Na Secção 2.2 é introduzido o conceito de ML, como os tipos de algoritmos que existem, os tipos de análises e é explicado o funcionamento de alguns algoritmos. Na Secção 2.3 são demonstradas algumas soluções já existentes de aplicações de ML em áreas relacionadas com este trabalho, nomeadamente previsão dos dados meteorológicos, deteção de fugas e uso sustentável da água, sendo ainda apresentado como estes podem contribuir para o desenvolvimento do trabalho, bem como este irá ser inovador.

2.1. Internet of Things

IoT pode ser denominada por uma rede de objetos físicos, tais como veículos, edifícios, equipamentos de construção, dispositivos de monitorização de saúde, entre muitos outros. A origem deste conceito remonta a 1982, e a primeira máquina ligada à internet foi uma máquina de refrigerantes, esta apresentou algumas características únicas, tal como ser capaz de indicar o seu stock e se as bebidas mais recentes lá colocadas estavam ou não frias, conseguindo desta forma fazer a leitura da temperatura e relatar o estado da mesma [**1**]. O conceito de IoT foi proposto pela primeira vez no final do século XX, mais concretamente em 1999, por Kevin Ashton, sendo na altura referido como algo nunca antes visto, capaz de revolucionar o mundo tal como o conhecíamos, em relação à tecnologia [**2**]. IoT permite que os objetos físicos consigam sentir, executar e também pensar, comunicando entre si, para compartilhar informações e coordenar decisões. Um grande objetivo do IoT é conseguir fazer com que se reduza a necessidade de intervenção humana nas atividades diárias que apresentem ser mais simples [**3**]. Internet of Things está cada vez mais ligada à vida das pessoas, como é possível perceber pelo estudo realizado pela Statista [**4**], representado na Figura 2.1, em 2018 haviam 22 mil milhões de dispositivos IoT conectados, enquanto que este número crescerá para 38.6 mil milhões em 2025 e por fim, em 2030 serão

50 mil milhões. Portanto estamos perante um grande aumento destes dispositivos ao longo dos próximos anos.

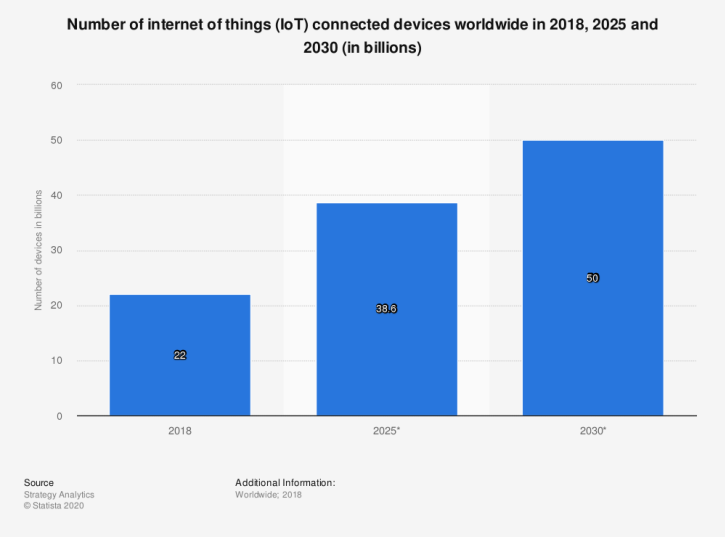

FIGURA 2.1. Número de Dispositivos Conectados por IoT em 2018, 2025 e 2030

Este aumento de dispositivos leva a que o IoT esteja cada vez mais presente no dia a dia das pessoas, mas não é só através de dispositivos que o IoT consegue atuar [**5**]. Tal como representado na Figura 2.2, um sistema de IoT é composto por seis elementos, que se interligam entre si para permitir que as soluções de IoT consigam recolher informação, transmiti-la, analisá-la e por fim decidir e informar o utilizador sobre o que pode ser feito.

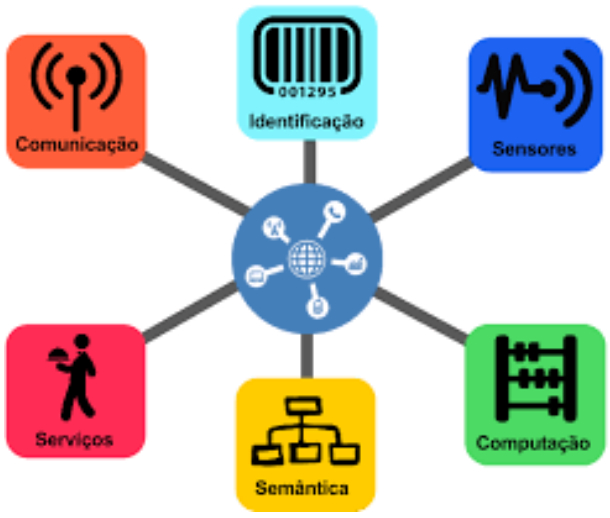

FIGURA 2.2. Elementos de IoT

Comunicação: Através da existência de comunicação pode-se trocar dados e também realizar serviços. Normalmente a comunicação mais usada entre os dispositivos é a comunicação sem fios, sendo esta otimizada para consumir o menor valor de energia possível.

Com os avanços da tecnologia, os protocolos mais usados tendem a variar ao longo dos anos, sendo neste momento os mais utilizados o Bluetooth, Wireless Fidelity (Wi-Fi) e Long-Term Evolution (LTE) e Long Range Wide Area Network (LoRaWAN).

Identificação: Tem um papel fundamental, uma vez que permite a distinção de diferentes dispositivos, de modo a existir combinação de serviços entre os mesmos. Tanto o Protocolo de Internet (IP) versão 4 como a versão 6 estão contemplados nos métodos de endereçamento de objetos.

Computação: A computação pode ser denominada como sendo o cérebro do IoT pois é a computação que permite total coordenação para o funcionamento, usando para isso tanto hardware, principalmente microcontroladores como Arduino, Raspberry Pi ou ESP32, bem como plataformas de software cloud, tais como Amazon AWS, Microsoft Azure ou IBM Watson.

Semântica: É a capacidade de agrupar todo o conhecimento, extraindo este conhecimento inteligente por meio de máquinas, para facultar os serviços ao sistema. Na extração considera-se a descoberta e a análise da informação para ser possível escolher a melhor opção tendo em conta o serviço desejado.

Serviços: Normalmente dividem-se em quatro diferentes níveis, serviços ubíquos, serviços relacionados com a identidade, reconhecimento de colaboração e por fim, de adjunção de informações.

Todas estas funcionalidades permitem ao IoT ter a capacidade de se adaptar a qualquer situação ou especificação, permitindo assim ter múltiplas aplicações, sendo já adotada por vários setores, desde a saúde, transportes, retalho, construção e agricultura [**6**]. E com esta vasta aplicação é percetível o aumento do uso de IoT e da previsão do aumento do uso do mesmo no futuro.

2.1.1. Internet of Things na Agricultura

IoT na agricultura ou agricultura de precisão é a utilização de tecnologia avançada para que se consiga recolher dados, processar esses dados, tudo de forma inteligente e através de dispositivos que estejam ligados à internet. O conceito de IoT pode ser cada vez mais encontrado na área da agricultura, principalmente por ser uma área que apresenta um elevado gasto de recursos hídricos, e também com o objetivo de tornar a agricultura mais rentável em todos os seus sentidos, obtendo no final melhores colheitas e de forma mais económica.

No artigo [**7**] foi instalada uma Rede de Sensores sem Fios (RSSF) que melhorou a eficácia e também a eficiência dos agricultores. A rede de sensores sem fios serve para monitorizar e fazer o controlo de fatores que influenciam toda a produtividade da cultura em questão e também o seu crescimento, tendo sido desta forma também possível controlar as variáveis, tais como, temperatura, estado do solo, humidade, entre outros. É assim possível verificar que através da instalação de uma simples rede de sensores sem fios se consegue atingir um maior controlo sobre todo o campo agrícola, os seus constituintes e também levar a uma redução de custos, sendo um tópico muito importante para a maioria dos agricultores. Neste artigo, os autores conseguiram que, ao regar em alturas estratégicas e, mantendo uma temperatura ideal entre 29ºC a 32ºC e a humidade entre 72-81%, o agricultor conseguisse ter vegetais e limões prontos em 2 meses depois de plantá-los, enquanto que sem este controlo, proporcionado pelo IoT, não seria possível saber ao fim de quantos meses os agricultores podiam ter a plantação pronta para apanhar, ajudando os mesmos a conseguir planear ao pormenor os tempos necessários para cada colheita, bem como a manter sempre as melhores condições possíveis. Por isso, mecanismos de IoT trazem vantagens na agricultura, nomeadamente por permitirem uma redução de custos, maior rentabilidade dos agricultores, através de automatização e monitorização constante, sem assistência permanente dos agricultores de uma forma convencional, melhorando também a sustentabilidade, segurança alimentar, permitindo ainda ajudar a que se cumpra na proteção do meio ambiente [**8**].

2.2. Machine Learning

As abordagens de análise de dados apresentadas anteriormente são bastantes importantes para o correto funcionamento do IoT, e para que estas funcionem de maneira automática e inteligente, machine learning é a ferramenta ideal para realizar toda esta análise. O campo científico de machine learning é um ramo da Inteligência Artificial (IA) que fornece aos sistemas a capacidade de aprender e melhorar automaticamente com a experiência, sem estes serem explicitamente programados. Enquanto que a IA educa o sistema a fazer determinadas coisas, o machine learning concentra-se no desenvolvimento de programas de computador que podem utilizar dados para aprender de forma autónoma. O processo de aprendizagem começa com observações ou dados, como por exemplo, experiência direta ou instrução, a fim de procurar padrões nos dados e ser capaz de tomar melhores decisões no futuro com base nos exemplos fornecidos. O objetivo é que o computador possa então aprender automaticamente sem haver intervenção humana [**9**]. Sem

o uso de machine learning, a análise de dados é bastante limitada e não está apta para lidar com mudanças rápidas de dados e dados não estruturados [**10**].

2.2.1. Tipos de Algoritmos

Em relação aos tipos de algoritmos, estes são categorizados em quatro tipos, sendo eles: aprendizagem supervisionada, semi-supervisionada, não supervisionada e por reforço [**9**].

Supervisionados: Nestes algoritmos há uma primeira fase em que se ensina ou se treina a máquina usando dados que estão bem rotulados, consistindo em pares de entradas e saídas. O objetivo é que a máquina consiga aprender uma regra geral que relacione as entradas com as saídas. Depois quando chega um conjunto de dados à máquina, o algoritmo analisa tendo em conta o conhecimento anterior que adquiriu a partir do treino e produz o resultado final [**9, 11**].

Não Supervisionados: Nestes algoritmos, consiste-se em treinar a máquina utilizando informações que não são classificadas nem rotuladas, permitindo que o algoritmo atue sobre essas informações sem orientação. A tarefa da máquina passa por agrupar informações não classificadas de acordo com semelhanças, padrões e diferenças, sem qualquer treino prévio de dados, tentando encontrar a "resposta certa" [**9, 12**].

Semi-Supervisionados: Em algoritmos deste tipo, são possíveis de se observar características dos algoritmos supervisionados e algoritmos não supervisionados. Os dados de treino são compostos tanto por dados rotulados como também por dados não rotulados, normalmente uma pequena quantidade de dados rotulados com uma grande quantidade de dados não rotulados. Ao existir um uso inteligente dos dados, poderá levar a uma melhoria de desempenho do modelo [**9**].

Por Reforço: Com estes algoritmos, a máquina é estimulada a descobrir através de testes do tipo "tentativa e erro" quais ações terão para a máquina uma maior recompensa. Com esta abordagem é possível ensinar um sistema a priorizar hábitos em detrimento de outros, tomando a melhor decisão diante de diferentes situações [**9**].

Na Tabela 2.1 é possível observar as vantagens e desvantagens dos diferentes tipos de algoritmos.

Tabela 2.1. Tipos de Algoritmos: Vantagens e Desvantagens

Tipos de Algoritmos	Vantagens	Desvantagens
Aprendizagem Supervisionada	- Permite guardar dados e produzir dados das experiências anteriores - Ajuda a otimizar os critérios de desempenho com a ajuda da experiência. - Ajuda a resolver vários tipos de problemas de computação do mundo real.	- Classificar Big Data pode ser um desafio na medida em que pode ser complexo. - O treino exige muito tempo.
Aprendizagem Não Supervisionada	- Mais independente, não precisa de treino prévio	- Não existe referência para avaliar precisão do modelo
Aprendizagem Semi Supervisionada	- Não é necessário treinar os dados todos previamente. - Uso inteligente dos dados	- Pode existir pouca referência para avaliar precisão do modelo
Aprendizagem por Reforço	- Completamente independente, descobre por si só, através de recompensas	- Pode demorar algum tempo até perceber o que é "certo" ou "errado"

2.2.2. Tipos de Análises

Dentro dos vários tipos de aprendizagem que existem, cada um pode ser dividido em vários tipos de análises, em que os mais conhecidos são a análise por classificação, regressão e por clustering.

Classificação: A classificação está inserida na aprendizagem supervisionada. Consiste no processo de atribuir um rótulo a um tipo de entrada de dados. Geralmente pode ser mais utilizada esta abordagem em casos em que as previsões são de naturezas distintas, por exemplo, sim ou não. De uma maneira mais científica, o que foi dito pode ser representado na Equação 2.1, tal como o autor identifica, *"is a predictive model that approximates a mapping function (f) from input variables (x) to identify discrete output variables (y)"* [**13**].

$$y(f : x- > y) \tag{2.1}$$

Regressão: É também uma subcategoria de aprendizagem supervisionada, mas geralmente usada em casos que os dados previstos diferem de sim ou não. Pode existir regressão, podendo ser do tipo linear simples ou então do tipo linear múltipla, como o nome indica, na linear simples estamos perante apenas uma variável independente, enquanto que na regressão linear múltipla, esta refere-se a várias variáveis independentes.

Clustering: Este tipo de análise está inserido no modelo de aprendizado não supervisionado, é geralmente usado para encontrar agrupamentos naturais de dados, este agrupamento denomina-se por *clusters* [**14**], sendo estes agrupamentos definidos segundo o seu grau de semelhança. Um possível exemplo poderá ser para agrupar diferentes textos que tenham no seu conteúdo informações parecidas sobre o mesmo assunto e separar textos de conteúdos diferentes.

2.2.3. Algoritmos

Baseado no tipo de aprendizagem e de análise, existe uma grande variedade de algoritmos que podem ser implementados para obter os resultados desejados. Dada a vasta quantidade de algoritmos, e sendo impossível apresentá-los e estudá-los a todos, foram tidos em consideração os que poderão ser mais importantes do ponto de vista da precisão para o foco deste trabalho, tendo o cuidado de procurar pelo menos um algoritmo de cada tipo de aprendizagem.

2.2.3.1. *Árvore de Decisão (AD).* O algoritmo árvore de decisão representa o modelo de classificação de um conjunto de dados na forma de estrutura de árvore, dividindo o conjunto de dados em subconjuntos menores. O nó raiz (nó principal) é escolhido tendo como base o ganho de informação de atributos num conjunto de dados. Estes dados sofrem uma aplicação chamada de mecanismo de cálculo de entropia, este é aplicado de forma a ser possível calcular o ganho de informação [**15**]. Desta maneira consegue-se usando apenas regras pequenas de decisão chegar ao modelo final de decisão. Pode-se observar o funcionamento do algoritmo na Figura 2.3.

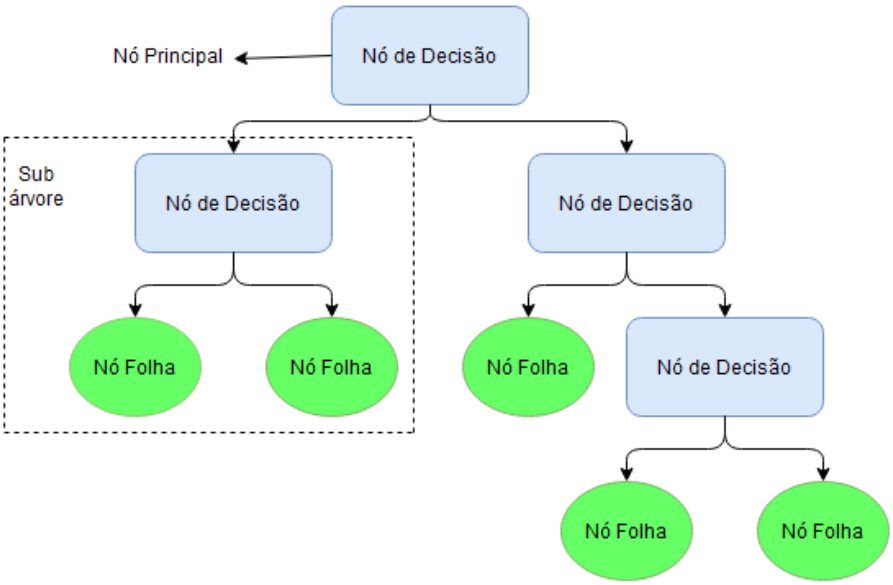

FIGURA 2.3. Arquitetura Árvore de Decisão

2.2.3.2. ***Máquina de Vetores de Suporte (SVM).*** O SVM é um classificador binário que constrói um hiperplano de separação linear com o fim de classificar instâncias de dados. Os recursos de classificação de SVMs tradicionais, usando o "truque do kernel" [**14, 16**], conseguem ser aperfeiçoados por meio de transformação do espaço de recursos original para um espaço de recursos de uma dimensão superior. Na Figura 2.4 pode-se observar a arquitetura do algoritmo.

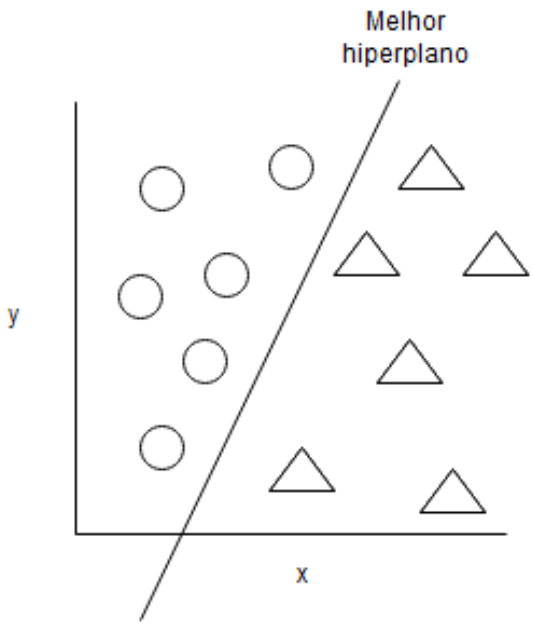

FIGURA 2.4. Arquitetura SVM

Um conjunto de dados de treino D pode ser representado na Equação 2.2, onde xi representa um dado de entrada e tem consequentemente um valor alvo yi. A função de SVM pode ser descrita como na Equação 2.3.

$$D = (x_1, y_1), ..., (x_n, y_n) \tag{2.2}$$

$$f(n) = (w \times \Phi(x)) + b \tag{2.3}$$

onde Φ significa uma transformação não linear de Rn para um elevado espaço dimensional, por sua vez $w \subset R^n$ e $b \subset R$. É necessário conseguir satisfazer a condição de Mercer, esta condição corresponde ao produto interno dum espaço de recurso [**17**]. SVMs podem ser usados nos seguintes tipos de análise: classificação, regressão e clustering [**14**].

2.2.3.3. ***Rede Neuronal Artificial (RNA).*** As Redes Neuronais Artificiais (RNA) têm como base os processos biológicos que existem no cérebro humano. Tem de existir um treinamento para ser possível prever padrões semelhantes em dados futuros. Uma das grandes vantagens das redes neuronais artificias é serem capazes de adotar a sua complexidade sem mesmo antes conhecerem os princípios subjacentes [**18**]. A arquitetura pode ser vista na Figura 2.5.

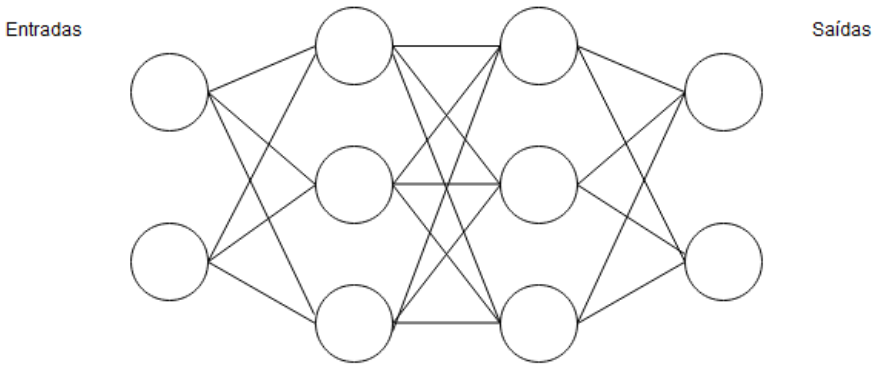

FIGURA 2.5. Arquitetura Rede Neuronal Artificial

2.2.3.4. ***Floresta Aleatória (FA).*** Este algoritmo encontra-se nos algoritmos de aprendizagem supervisionada, pode ser aplicado tanto para problemas de classificação como para regressão, em que é criada uma combinação de árvores de decisão, cada árvore dá uma classificação para que se alcance uma melhor precisão final. O algoritmo pode-se dividir em duas partes, na primeira parte pode-se ver que se trata da criação de um classificador de floresta e a segunda parte, que é depois de haver a premonição de resultados.

Uma das suas vantagens de utilização é que é um algoritmo simples de entender, de fácil implementação e pode ser usado mesmo para conjuntos que tenham bastantes dados e inclusive, que sejam universos de dados diferentes. O Funcionamento do algoritmo pode ser observado na Figura 2.6.

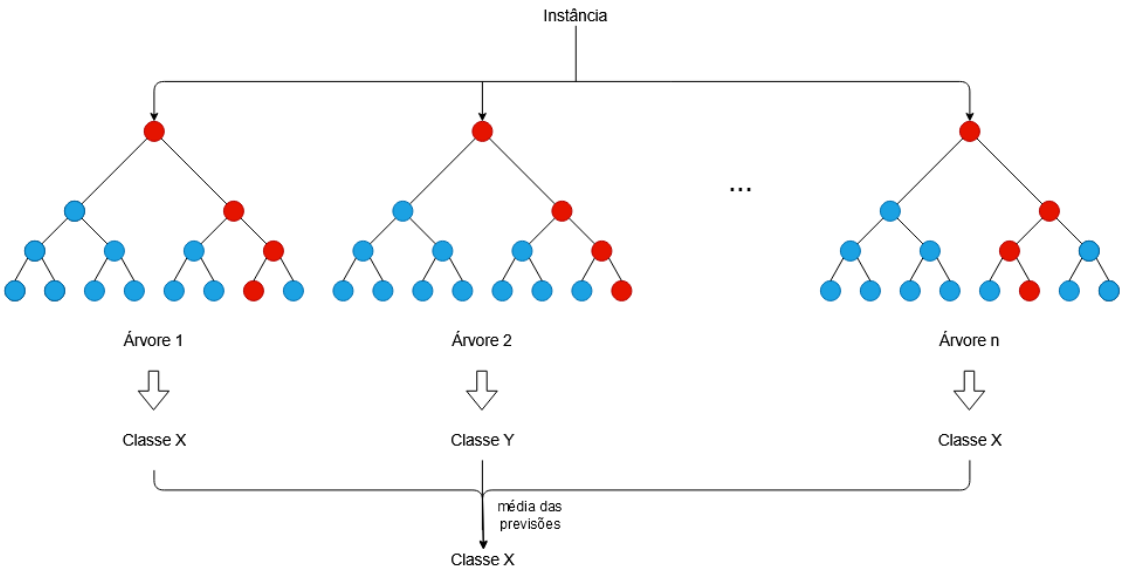

FIGURA 2.6. Arquitetura Floresta Aleatória

2.2.3.5. ***Extreme Gradient Boosting (XGBoost).*** É um algoritmo que pode ser usado para análise de regressão, como para classificação. Tem como base na sua implementação o algoritmo de árvores de decisão, tornando essa implementação mais precisa através de uma abordagem interativa com técnicas de ensemble [**19**]. O algoritmo constrói o modelo faseadamente, de uma forma genérica e através disso consegue otimizar uma função de perda diferenciável arbitrária. O algoritmo cria vários modelos, através de uma aproximação obtêm-se o valor chamado de "resíduo", através deste valor é possível obter a distância que foi prevista pelo algoritmo em comparação com o valor real. Depois de se obter o valor do resíduo, o segundo modelo já tem como base esse valor e volta a ser extraído um novo valor de resíduo. As somas dos ajustes de todos os modelos calculados anteriormente dão origem ao modelo final [**20, 21**].

2.2.3.6. ***K-Means Clustering.*** Este algoritmo é de aprendizagem não supervisionada e começa por selecionar um grupo de pontos para cada cluster, de seguida procede à realização de cálculos iterativos de forma a obter a melhor posição do ponto central, este passo repete-se até que que por fim todos os pontos convergirem e não houver mais alterações ao nível dos centros dos clusters. Cada cluster terá pontos semelhantes entre

si. O parâmetro k representa a o número de clusters que vão existir no algoritmo e é necessário defini-lo no início do processo. Na Figura 2.7 pode ser visto um exemplo em como é antes e depois de se aplicar o algoritmo.

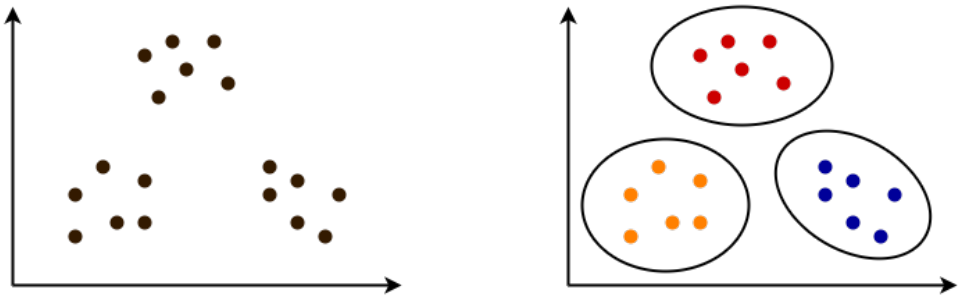

FIGURA 2.7. Antes e Depois de se aplicar o algoritmo K-Means Clustering

2.2.3.7. ***Máquina de Vetores de Suporte Semi-Supervisionada (S3VM).*** Este algoritmo tem um funcionamento muito similar ao que foi visto quando foi apresentado o conceito de SVM, no entanto é um algoritmo de aprendizagem semi-supervisionada, foi proposto primeiro por Bennet e Demiriz, e este baseia-se na suposição de cluster. O algoritmo tem como objetivo, usar dados que sejam rotulados previamente e dados que não sejam rotulados de forma a construir um classificador como se poderá ver na imagem seguinte. Existem algumas fórmulas mais teóricas e podem ser consultadas mais ao pormenor no trabalho [**22**]. Na Figura 2.8 pode-se observar um exemplo da utilidade deste algoritmo, em que nas linhas tracejadas está o exemplo ao usar-se o algoritmo SVM, mas ao serem introduzidos dados não rotulados o algoritmo SVM já não conseguiria lidar com o exemplo e o algoritmo Máquina de Vetores de Suporte Semi-Supervisionada (S3VM) conseguiu separar melhor os dados.

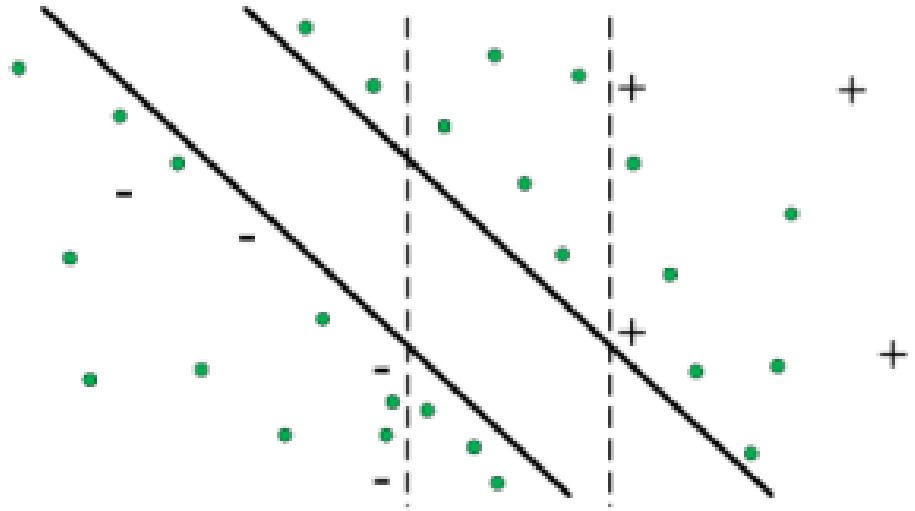

FIGURA 2.8. Exemplo usando o algoritmo S3VM

2.2.3.8. *Q-Learning.* Este algoritmo é de aprendizagem por reforço e ao contrário da maior parte dos algoritmos aqui apresentados, não segue um modelo de forma a aprender quais as ações que deve escolher consoante as circunstâncias, pode ser usado em situações que existam recompensas ou transições estocásticas, sem necessitar de adaptações prévias. O algoritmo procura sempre obter a melhor ação, dado o estado atual de modo a maximizar a recompensa, tem como base resolver problemas que sejam complexos, podendo ser usado quando o problema original é tão complexo que se divide em várias partes e é aplicado um algoritmo Q-Learning a cada uma das parte [**23**].

ML tem um papel fundamental porque tem a capacidade de resolver problemas complicados de uma forma escalonável. Com o aumento do uso de IoT e de ML, mais tipos de sistemas serão operados por estas tecnologias, e consequentemente será possível que se consiga reduzir os custos, aumentar a produtividade e a eficiência [**10**].

2.3. Aplicações com Machine Learning

Com o elevado número de alternativas para implementar sistemas de machine learning, é normal que existam já alguns estudos relativos à sua utilização nas mais diversas áreas. Na área da agricultura de precisão existem já vários trabalhos, tanto académicos como industriais que apresentam resultados positivos, mas que podem ainda ser melhorados. Sendo o sistema baseado na previsão meteorológica associada à análise de dados em tempo real vindos de sensores, de modo a identificar como melhorar o uso de água para que este seja feito de uma forma inteligente e sustentável, foram identificados alguns trabalhos que abordam estes temas, de modo a conseguir identificar os seus pontos fortes

e fracos, garantir uma base de partida sólida e comparação dos mesmos para o sistema a desenvolver.

As aplicações com base em ML são várias e nos últimos anos começou a haver alguma investigação e implementação desta tecnologia focada na agricultura, no entanto é algo ainda recente e ainda existe espaço para melhorias. Na maior parte dos trabalhos foram utilizados algoritmos de aprendizagem supervisionada visto serem os algoritmos que apresentaram melhores resultados ao longo dos últimos anos de investigação, no entanto neste trabalho vão ser abordados algoritmos de cada tipo de aprendizagem.

No trabalho [**19**], o autor apresenta uma solução baseada em tecnologia IoT e ML para controlo automático de rega, usando nomeadamente uma rede de sensores e atuadores wireless. Neste trabalho o autor implementou vários algoritmos de ML, tendo concluído que o algoritmo com melhor valor de precisão foi o XGBoost, que obteve 87.73%. Na aplicação do sistema desenvolvido foi possível uma poupança de 60% de água. Este trabalho de dissertação será uma continuação do trabalho [**19**], e tentará aumentar este valor de precisão, bem como o valor de poupança de água, focando-se mais na parte de ML. Em termos de desvantagens deste trabalho não foi tido em conta os outros tipos de aprendizagem, não podendo fazer comparações se de facto com aprendizagem supervisionada se obtém a melhor precisão possível e o autor preocupou-se apenas em descobrir a melhor hora para regar. De forma a obter melhores resultados e introduzir inovação no sistema, a nossa proposta, além de obter a melhor hora de rega, irá também incluir outros fatores, como a necessidade efetiva de rega se chover nas próximas horas ou se existir alguma fuga, entre outros.

No trabalho [**24**], os autores realizaram um estudo comparativo de previsão do tempo usando algoritmos de ML, tais como, Rede Neuronal Artificial, Regressão Linear Múltipla, Máquina de Vetor de Suporte e também modelos híbridos. Como resultado do trabalho foi possível observar que o modelo que se mostrou mais preciso foi o modelo de Regressão Linear Múltipla com um erro médio de 1,0782, e de um valor de 0,8119 de correlação entre a temperatura que efetivamente esteve com a temperatura calculada pelo algoritmo [**24**]. Quanto menor o erro melhor o algoritmo e quanto mais perto o valor de correlação estiver de 1, melhor a relação entre as duas variáveis, portanto ainda se conseguirá melhorar estes valores. Também de salientar o trabalho [**25**], em que os autores utilizaram um modelo neuronal híbrido de Multilayer Perception e Radial basis function para previsão em tempo real com parâmetros de entrada com valores passados naquela região tais como

temperatura mínima, precipitação, velocidade do vento, entre outros. Para uma previsão de dois de antecedência o modelo mostrou-se ser eficaz, apresentando o valor de 0.93 para o coeficiente de correlação.

No trabalho [**17**], os autores estudaram que através de Machine Learning é possível usar a água de uma forma mais sustentável. Estes fizeram demonstrações de alguns algoritmos de ML e chegaram à conclusão que os mesmos podem ser vantajosos para o planeamento e gestão de recursos hídricos. Concluíram que as técnicas de ML têm um grande potencial quando usadas para este fim, uma vez que estes são capazes de fazer modelos robustos, sendo eficientes para a área ligada à sustentabilidade da água. Os algoritmos usados foram a Rede Neuronal Artificial, Fuzzy Logic, sistema neuro fuzzy adaptativo de inferência, análise do componente principal, máquina de vetores de suporte e o M5 Model Tree and Reduced-Error Pruning Tree, tendo obtido para o melhor valor de erro percentual médio 18.22% e 36.63% para o conjunto de treino e de validação respetivamente.

CAPíTULO 3

Arquitetura do Sistema

Esta dissertação tem por objetivo desenvolver um sistema suportado por Machine Learning capaz de identificar a necessidade de rega de um campo agrícola baseado na sua localização, condições do solo, tipo de colheita e das condições meteorológicas. Para atingir este objetivo será necessário desenvolver um algoritmo capaz de obter os valores de vários parâmetros para um determinado dia e local, parâmetros esses tais como a temperatura, precipitação, evapotranspiração, humidade relativa, entre outros, que serão depois analisados de modo a indicar se é necessário regar, qual o melhor período do dia para o fazer e a duração necessária. Será tido em conta o que foi estudado no capítulo anterior, servindo como base e aprendizagem para esta dissertação ser realizada com sucesso e cumprir os objetivos. A arquitetura do sistema a desenvolver é composta por um conjunto de módulos, que quando associados permitem atingir o objetivo traçado. Esses módulos são compostos por:

- Condições meteorológicas: Obtenção das condições meteorológicas do último dia e das próximas 4 horas;
- Condições do solo: Obtenção das condições atuais do solo e previsão das mesmas nas próximas 4 horas;
- Análise de dados: Análise dos dados obtidos de modo a decidir a necessidade e em caso favorável, a duração de rega;
- Apresentação de resultados: Informar o utilizador sobre a decisão.

3.1. Metodologia

Nesta secção serão explicadas de forma resumida as funções do sistema, desde o pedido por parte do cliente até ao resultado final. A Figura 3.1 mostra o funcionamento do sistema:

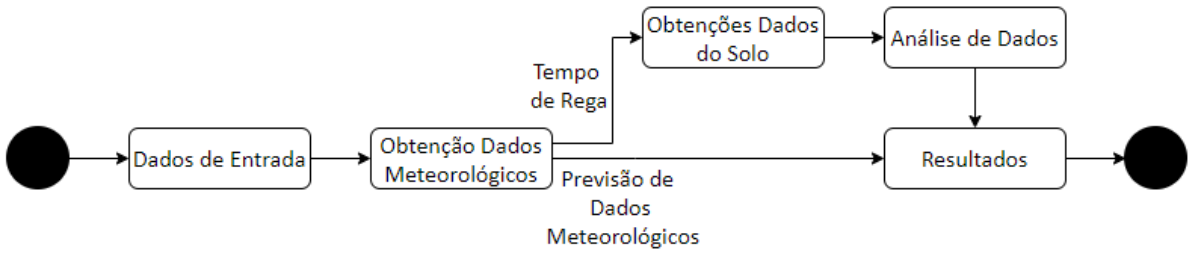

FIGURA 3.1. Funcionalidade do Sistema

O sistema inicia a sua funcionalidade recebendo um conjunto de parâmetros por parte do utilizador, nomeadamente a localização, dia para análise e indicação se pretende apenas receber as informações meteorológicas ou a avaliação para o sistema de rega. Caso o utilizador escolha a segunda opção então o sistema também receberá por parte do utilizador os dados referentes ao tipo de rega e às características do campo.

Independentemente da decisão do utilizador, o sistema começa por calcular as condições meteorológicas recorrendo à previsão dos dados meteorológicos, através de modelos de machine learning previamente treinados e guardados, para a data indicada pelo utilizador. Com essas previsões, o sistema calcula ainda os valores da precipitação das últimas 24 horas e faz a previsão dos dados meteorológicos para as próximas 4 horas.

Caso o utilizador tenha optado apenas por receber as informações meteorológicas, o sistema termina a sua funcionalidade devolvendo ao utilizador as informações obtidas.

Caso o utilizador tenha optado pela decisão sobre os períodos de rega, o sistema irá obter as condições atuais do solo e previsão para as próximas 4 horas, através de um algoritmo de machine learning, baseado nas especificações do campo agrícola fornecidos pelo utilizador.

No passo seguinte, é feita uma análise aos dados, começando por se verificar se os valores obtidos para a condição do solo são favoráveis à rega. Caso não sejam, não é feita a análise para essa data, visto não serem cumpridas as condições impostas para uma rega sustentável e eficiente. Caso contrário, para as datas em que os valores obtidos cumpram as condições do solo, são usados um conjunto de equações matemáticas de forma a analisar se há necessidade de rega. Aquando da necessidade de rega, o sistema verifica se ao longo das próximas 4 horas existe alguma condição adversa à rega, escolhendo depois a data e hora em que se identifica o melhor período para uma rega sustentável e eficaz.

Na apresentação de resultados, informa-se o utilizador sobre a decisão gerada ao longo do sistema, em caso da decisão ser favorável para a rega, indica-se ao utilizador qual a melhor hora para regar e a duração da mesma.

As secções seguintes irão detalhar cada um dos passos apresentados na metodologia.

3.2. Extração de Conhecimento

Nesta secção serão demonstrados quais os tipos de extração de conhecimento que existem ao longo do sistema, sendo que este é referente à meteorologia, às condições do solo e para a decisão de rega.

Os dados de entrada são utilizados para se proceder às várias análises de dados de cada subsecção. Estas análises são necessárias para o sistema obter o objetivo de prever a necessidade de rega e a duração da mesma.

Os dados referentes à meteorologia e às condições do solo são bastante importantes ao longo do trabalho, uma vez que todo este vai girar em torno destes dados. Estes dados são previstos através de um algoritmo de machine learning que é treinado utilizando um dataset, sendo que com estes dados previstos será possível apresentá-los ao utilizador ou usá-los para posteriormente ser feita uma análise e poder ser indicado o tempo de rega.

As condições são analisadas de forma a se perceber se existe necessidade ou não da rega, caso os valores não sejam concordantes com as imposições derivadas das condições do solo, estamos perante casos que não faz sentido prosseguir com a sua análise, os restantes valores, que são datas favoráveis para a rega, prosseguem para a próxima análise.

3.3. Análise de Dados

De forma a ser possível decidir a necessidade de rega, são tidos em conta os dados provenientes da secção anteriormente descrita e com estes dados é possível perceber quais as datas que são favoráveis à prática da rega. São de seguida calculadas, através de equações matemáticas, quanto tempo é necessário regar para cada data, sendo que destas datas se analisa qual a data que apresenta o resultado mais favorável. É ainda necessário verificar o que acontece em cada uma das datas posteriores, porque podem, por exemplo, existir períodos de chuva, tornando essa data e as anteriores desfavoráveis à prática de rega sustentável. Nestas situações torna-se necessário analisar as datas posteriores para obter o melhor período para regar. Após esta análise consegue-se por fim saber se existe ou não a utilidade de rega.

Sabendo os valores dos dados meteorológicos e das condições do solo, tais como a temperatura, velocidade do vento, precipitação, evapotranspiração, humidade relativa e humidade do solo, é possível criar um conjunto de regras com a finalidade das mesmas servirem como decisão para demonstrar se é favorável ou não a rega naquela data.

3.4. Plataformas de Suporte

De forma a conseguir alcançar os objetivos nesta tese foi necessário criar uma aplicação web de modo a oferecer ao utilizador uma plataforma interativa onde este possa introduzir as especificações do campo agrícola para análise e receber as informações sobre as previsões meteorológicas e os tempos de rega.

Para isto, recorreu-se à linguagem de programação Python e à Framework Flask, uma framework que permite criar de uma forma simples uma aplicaçao web capaz de interagir com o utilizador, sem a necessidade de bibliotecas ou ferramentas externas. Recorreu-se ainda a vários templates, com base em html, para construir as páginas que o utilizador terá acesso. Esta aplicação web, Figura 3.2, permite pedir ao utilizador, através de um formulário, introduzir os valores e retornar na mesma página, ou noutra, a resposta com os dados pretendidos.

Para implementar toda a parte interna do código, foi criado um script Python que permite receber várias entradas que o utilizador preenche, fazer toda a parte da análise de dados que foi descrita anteriormente, incluindo saber as previsões dos dados meteorológicos e/ou o tempo necessário que precisa de regar o campo, e devolver esses resultados para serem apresentados ao utilizador.

FIGURA 3.2. Menu da Aplicação Web

Como se pode observar, existem 3 opções, a primeira opção serve para voltar para esta mesma página, existe depois outra opção, sendo a opção para o utilizador consultar a previsão para os dados meteorológicos futuros e na terceira opção o utilizador é redirecionado para a página referente à previsão da duração do tempo de rega.

Em relação à aplicação web, também se resolveu criar uma Interface de Programação de Aplicações (API), um endpoint web sem formato interativo e visual, para dar oportunidade do utilizador fazer o pedido da informação através de dispositivos, usando pacotes HTTP onde a informação é passada através de pedidos no URL.

Foi criado um script Python, que à semelhança do anterior recebe os vários valores para a análise. Neste cenário os valores são fornecidos não por um formulário, mas sim por um pedido GET num pacote HTTP, sendo, depois da análise pretendida, devolvido um JSON com o resultado obtido. Através desta funcionalidade é igualmente possível o utilizador escolher os valores que quer colocar nos vários parâmetros e obter os valores dos dados meteorológicos ou os tempos de rega.

Independentemente da maneira como se interage com a plataforma de suporte, o resultado será o mesmo uma vez que o script de decisão usado é igual em todas as opções. Também em ambos os casos existe uma validação dos parâmetros fornecidos pelo utilizador de forma a garantir que o utilizador, por exemplo, só consiga preencher os campos numéricos apenas com números.

CAPíTULO 4

Modelo de Análise de Dados - Condições Meteorológicas

Como descrito na metodologia do Capitulo 3 é necessário obter as condições meteorológicas que irão estar numa determinada data e local. Neste capítulo o objetivo é fazer uma investigação em relação ao algoritmo de machine learning que é mais eficaz para a previsão das condições meteorológicas. Esta investigação é importante porque depois de escolhido o modelo, este será usado no desenvolvimento de um algoritmo que implementa o modelo escolhido neste capítulo para decidir a necessidade de irrigação.

Torna-se cada vez mais importante conseguir-se alimentar a população que faz parte do nosso planeta, para isso é necessário ter maneiras de produzir alimentos e tendo em vista a maneira mais eficiente. Com o surgimento e a evolução da Internet of Things e de Machine Learning torna-se possível controlar e gerir todo o processo ligado à agricultura, sendo possível decidir tendo em conta factores como o tempo que esteve, que está e que irá estar, qual será a melhor altura para regar o campo agrícola. Algo bastante complexo para um agricultor conseguir decidir da forma mais eficiente.

A previsão do tempo consiste em ser possível prever o estado da atmosfera num local específico e num momento posterior ao atual [**26**]. A previsão antes de existirem os conceitos de IoT, ML e IA era feita com muito esforço dos humanos, passando numa segunda fase a funcionar com base na Previsão Numérica do Tempo (NWP) e só mais tarde é que começaram a aparecer as melhorias relacionadas com técnicas de Inteligência Artificial.

Vários estudos já demonstraram que IoT e ML podem ser aplicados com o uso de técnicas de aprendizagem para controlarem as doenças nas colheitas, verificar a qualidade dos alimentos, podendo ser usados para outros fins, mas um dos fins que destaca é a capacidade de conseguir prever os dados meteorológicos. Desta forma, sabendo o tempo que irá estar no futuro, pode ajudar a decidir se será melhor realizar-se a rega ou adiar a mesma.

Um sistema pioneiro que recorreu à inteligência artificial foi o Dynamical Integrated foreCast (DICast®), este sistema tenta fazer uma aproximação do que consiste a previsão

humana. Normalmente consegue uma melhoria em torno de 10 a 15% do que a previsão humana e, portanto, pode-se confirmar que a melhoria foi muito boa [**27**].

De forma a estudar este tema, foram estudados e usados através de regressão os seguintes algoritmos de aprendizagem supervisionada: Floresta Aleatória, Árvore de Decisão, Rede Neuronal e por fim, Regressão Linear. De seguida foi criada uma metodologia, tentando chegar ao melhor resultado possível.

4.1. Metodologia

A metodologia foi dividida em várias fases, numa primeira fase da metodologia, foi necessário obter um dataset, seguindo-se o pré-processamento dos dados. Nas fases seguintes realizaram-se vários testes aos algoritmos, analisaram-se os resultados e foram escolhidos os melhores algoritmos ao longo do capítulo. As várias fases podem ser observadas na Figura 4.1.

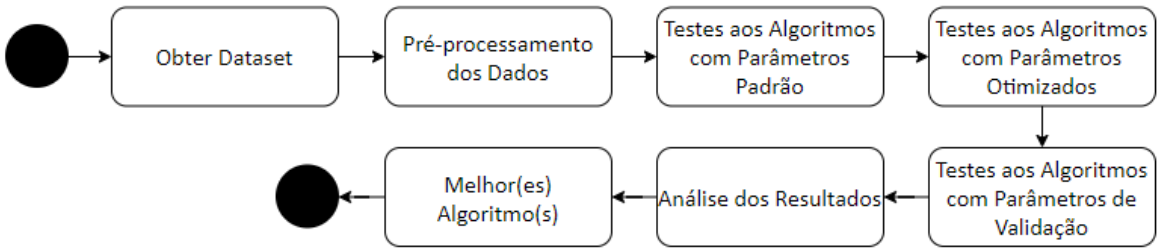

FIGURA 4.1. Fases da Metodologia

O dataset foi providenciado pela Agri4Cast database, dataset este composto com dados provenientes dos últimos 40 anos, dizendo os mesmos respeito à área do Alentejo, em Portugal, visto esta ser uma região predominantemente usada para a agricultura. O conjunto final de dados foi composto por 383526 entradas e podem-se observar todos os parâmetros que se encontram na Tabela 4.1.

Tabela 4.1. Propriedades do DataSet Diário

Parâmetro	**Descrição**
Latitude	Latitude do campo
Longitude	Longitude do campo
Altitude	Altitude do campo
Dia	Dia da observação
Mês	Mês da observação
Ano	Ano da Observação
Temperatura Máxima	Temperatura máxima [ºC]
Temperatura Mínima	Temperatura mínima [ºC]
Velocidade do Vento	Velocidade do vento[km/h]
Precipitação	Precipitação [mm]
Evapotranspiração	Evapotranspiração [mm/dia]

De forma a treinar o modelo de regressão, usou-se a linguagem de programação Python, as bibliotecas scikit-learn e o ambiente Anaconda.

Foi necessário proceder a um pré-processamento dos dados, nomeadamente na forma como a indicação do dia estava representada. Inicialmente o ano, mês e dia eram apresentados em apenas uma coluna, o que não ajudava a identificar os componentes específicos da mesma, por exemplo o mês, que tem uma importância grande quando tentamos correlacionar os dados de diferentes anos. Assim sendo, o dia, mês e ano de cada entrada foram colocados em três colunas independentes.

De forma a ser feita uma comparação rápida entre os algoritmos, treinou-se cada um deles com os parâmetros de configuração padrão, comparando a precisão obtida e o erro médio absoluto. De seguida utilizou-se uma configuração diferente, procedeu-se a uma hiper parametrização. Esta parametrização permite comparar várias configurações do algoritmo de forma a compreendermos qual a configuração que consegue alcançar os melhores resultados. Utilizou-se o método RandomizedSearchCV do scikit-learn, um método que permite ajustar e treinar os vários parâmetros de configuração dos algoritmos, decidindo de forma independente quais os parâmetros mais importantes para obter maiores benefícios [**15**].

Na última configuração, procedeu-se a uma validação cruzada, que permite provar que os valores são estáveis e se o modelo não está sub ou sobrestimado. Esta última configuração incrementa a certeza do desempenho do modelo.

4.2. Análise do Modelo e Observações

Como visto na secção da metodologia, foram realizados vários testes ao longo do processo e nesta secção serão observados e analisados os resultados que os mesmos obtiveram.

De forma a treinar, validar e testar cada modelo, dividiu-se a totalidade dos dados do dataset em três grupos, em que dos 100%, 70% foi utilizado para treino, 20% para validação e por fim, os restantes 10% para testes.

Os resultados obtidos para prever os parâmetros podem ser observados na Figura 4.2.

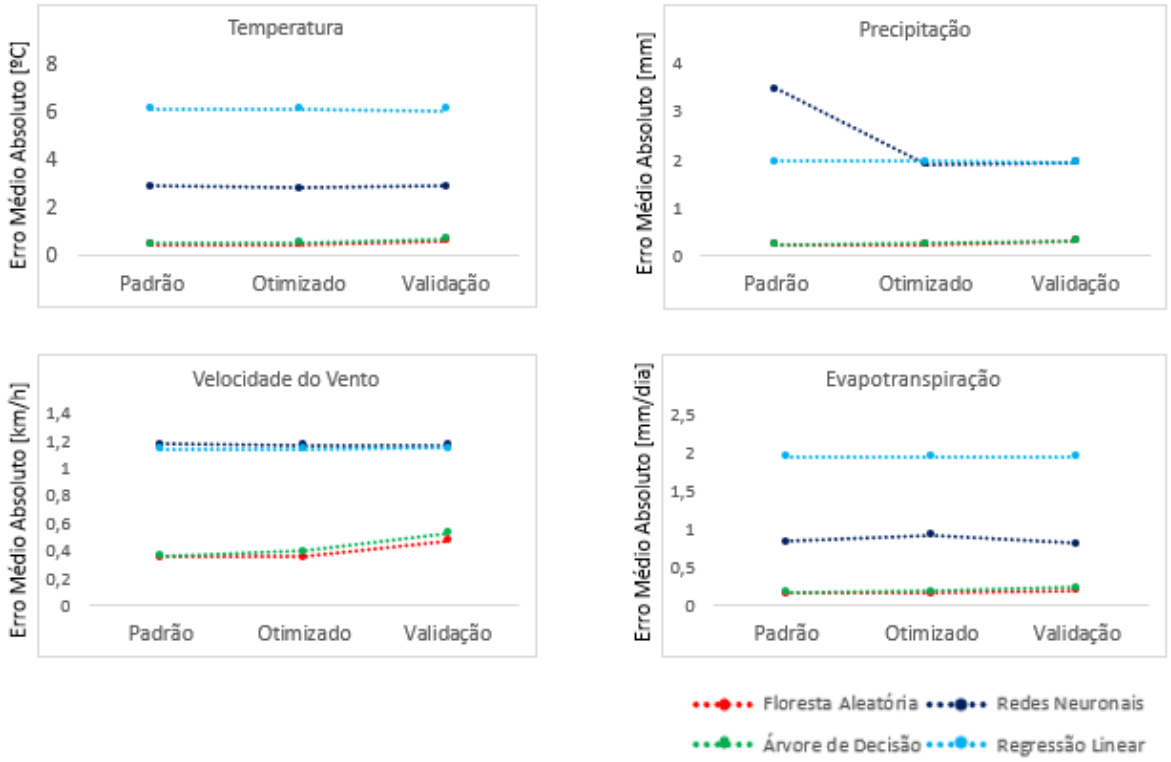

FIGURA 4.2. Resultados da Regressão com Dataset Diário

Pode-se concluir que os algoritmos floresta aleatória e árvore de decisão foram os que obtiveram melhores resultados, alcançando um valor de erro médio absoluto mais baixo. Floresta aleatória demonstrou obter melhores resultados para prever a temperatura, evapotranspiração e velocidade do vento com um erro médio absoluto de 0.639 ºC, 0.194 mm/dia e 0.472 km/h, respetivamente. Enquanto que o algoritmo Árvore de Decisão foi o melhor a prever a precipitação com um erro médio absoluto de 0.324 mm. Os restantes dois, uma vez que apresentaram os piores resultados em todos os cenários, foram excluídos dos próximos testes.

4.3. Dataset Diário ou Dataset Horário

Os resultados apresentados na secção anterior foram obtidos utilizando dados diários, usando a média dos valores obtidos durante aquele dia, dado ser esta a maneira mais comum que as plataformas meteorológicas emitem os seus boletins meteorológicos. Mas dada a constante alteração destes valores ao longo do dia e sendo que todas estas alterações podem ter um impacto na decisão dos momentos de rega, existe a necessidade de verificar se efetivamente usar dados diários ou dados horários pode ter alguma influência na precisão dos modelos de decisão.

Assim sendo, nesta subsecção procedeu-se à comparação de dois modelos de análise, um deles treinado com dados diários e outro com dados horários de forma a poder-se concluir qual deles será o mais indicado para se usar na previsão do tempo e que irá trazer mais vantagens na hora de decidir se é necessário regar ou por outro lado não. No que toca aos modelos utilizados, para o modelo diário foram utilizados os dados já apresentados na secção 4.2 que resultam nos resultados descritos na Figura 4.2.

Relativamente ao modelo horário, foi utilizado um dataset fornecido pelo MeteoBlue [**28**], composto com dados provenientes dos últimos 35 anos, também dizendo respeito à área Alentejo. O conjunto final de dados foi composto por 317736 entradas que inclui dados dos parâmetros apresentados na Tabela 4.2.

Tabela 4.2. Propriedades do Dataset Horário

Parâmetro	**Descrição**
Latitude	Latitude do campo
Longitude	Longitude do campo
Altitude	Altitude do campo
Hora	Hora da observação
Dia	Dia da observação
Mês	Mês da observação
Ano	Ano da observação
Temperatura Média	Temperatura média [ºC]
Humidade Relativa	Humidade relativa [%]
Humidade no Solo	Humidade no solo [%]
Velocidade do Vento	Velocidade do vento [km/h]
Precipitação	Precipitação [mm]
Evapotranspiração	Evapotranspiração [mm/dia]
Direção do Vento	Direção do vento [º]

Este modelo foi treinado utilizando a mesma metodologia descrita na secção 4.1 deste capítulo e os resultados obtidos podem ser observados na Figura 4.3, onde é feita a análise dos algoritmos com os dados do modelo horário. Neste caso foram apenas treinados o modelo floresta aleatória e árvore de decisão uma vez que foram os modelos que mostraram ser mais eficientes para calcular os parâmetros.

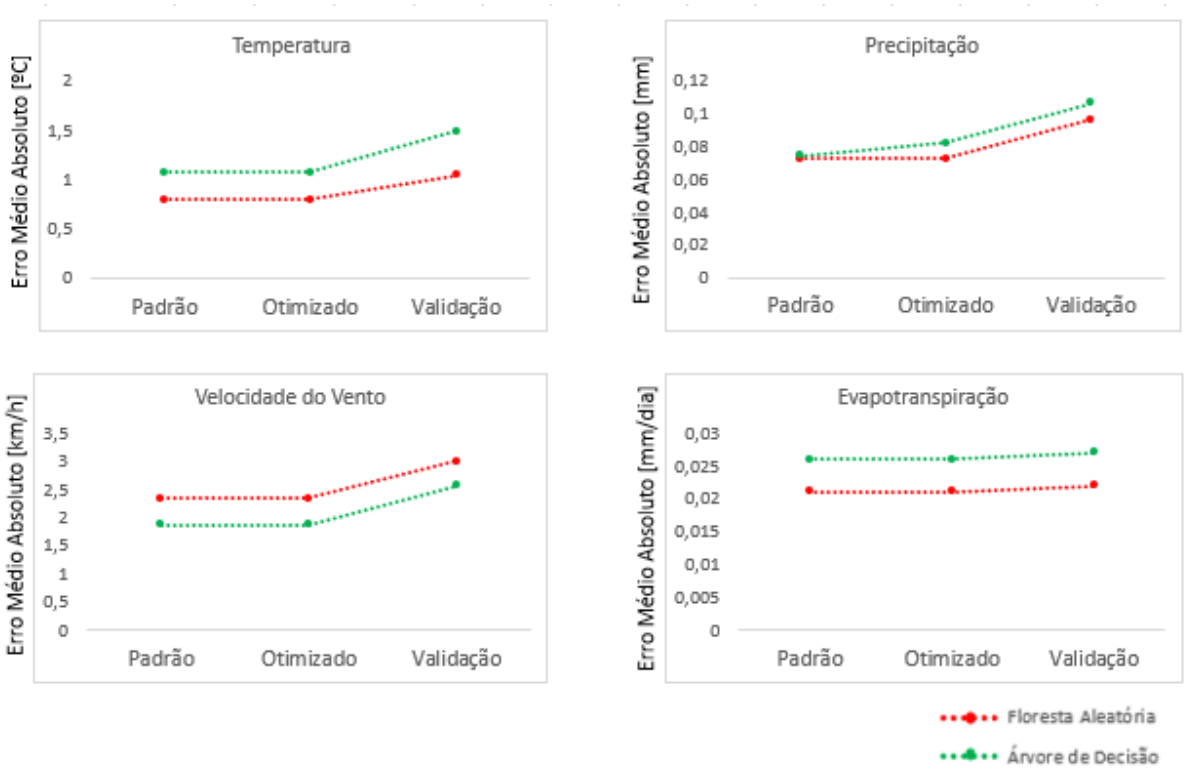

FIGURA 4.3. Resultados da Regressão com Dataset Horário

Comparando os parâmetros que são idênticos entre ambos os dataset, pode-se observar que nos resultados referentes à temperatura e à velocidade do vento com o uso de um dataset horário existe um aumento do erro médio absoluto, 1.043 ºC e 2.988 km/h respetivamente, quando comparado com o dataset diário, 0.639 ºC e 0.472 km/h respetivamente, em ambos com a utilização do modelo da floresta aleatória. Já para a precipitação os valores de erro médio absoluto obtidos com o dataset horário foram de apenas 0.096 mm enquanto que no dataset diário foi superior, 0.330 mm. Por fim para a evapotranspiração, obteve-se o valor 0.22 mm e com o dataset diário 0.194 mm.

Tendo em conta os resultados obtidos e as vantagens inerentes em usar o dataset horário, que permite que exista um universo maior de dados, dadas as 24 amostras diárias, torna-se evidente que, mesmo tendo uma precisão mais baixa em alguns parâmetros, a utilização do modelo horário é a escolha certa para o modelo. Em regra geral, para o

modelo árvore de decisão os resultados foram parecidos, no entanto como este mostrou-se ter uma eficácia menos boa, procedeu-se à utilização do modelo floresta aleatória.

Visto que o dataset horário oferece ainda valores relativos à humidade do solo e do ar, procedeu-se também à criação de um modelo de previsão para estes parâmetros, utilizando o algoritmo floresta aleatória, uma vez que este foi o que obteve melhores resultados nos testes ao longo deste capítulo.

Na metodologia de treino, apesar de só existir um modelo, resolveu-se fazer uma parametrização padrão, posteriormente fazer um ajuste de hiper parametrização e por fim obrigar o modelo a passar por uma fase de validação, de forma a averiguar que o modelo é estável e pode ser usado para prever valores futuros recebendo parâmetros de entrada tais como a data e local.

Os resultados obtidos dos restantes parâmetros podem ser observados na Figura 4.4.

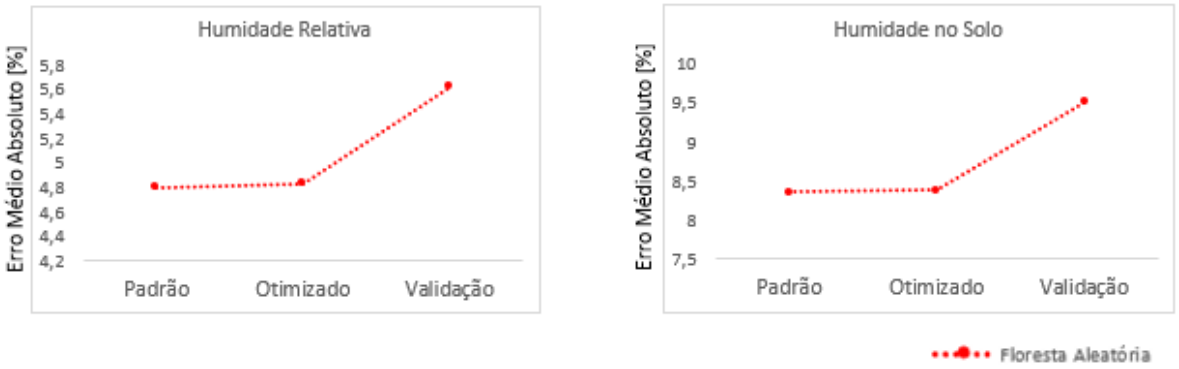

FIGURA 4.4. Resultados dos Novos Parâmetros da Regressão com Dataset Horário

Em suma, os valores de erro médio absoluto são aceitavelmente baixos, não tanto como esperado mas transmite que os modelos são eficazes e podem ser usados num cenário real.

4.4. Validação do Modelo

Após o treino dos vários modelos com os diferentes datasets e com a conclusão que o modelo horário apresenta valores menos precisos mas com a vantagem de dar valores mais detalhados ao longo do dia, ao invés do modelo diário que apenas dá um valor médio para esse dia, procedeu-se à exportação do modelo horário de modo a poder implementá-lo no algoritmo de decisão, utilizando a ferramenta joblib [**29**].

De modo a validar a eficiência do modelo criado, o mesmo foi aplicado de modo a prever as condições meteorológicas para a totalidade do ano de 2020. Para isso foi criado

um script onde o modelo foi importado e ao qual foi pedido que analisasse todas as datas e horas disponíveis no ano de 2020, baseado também numa posição fixa dada através da latitude, longitude e altitude. Apesar de presentes no dataset utilizado para treinar o modelo, as informações de 2020 não foram utilizadas no treino, garantindo assim que o modelo vai ser testado com dados nunca antes vistos e permitindo também comparar os resultados obtidos com os valores reais para as situações testadas.

Foi assim possível prever os vários parâmetros, simulando o que irá acontecer no sistema quando for necessário prever dados futuros e comparar com os valores reais obtidos pelas estações meteorológicas. Depois de calculados os valores, procedeu-se à sua comparação de modo a obter a diferença real entre o valor que o sensor leu com o valor previsto pelo modelo de ML. Desta forma é possível concluir se o erro real que existiu durante a previsão dos valores está de acordo com o erro teórico obtido aquando do treino do modelo e perceber se o mesmo é aceitável e se permite ao modelo de previsão substituir os valores reais fornecidos em tempo real pelas estações meteorológicas. A Tabela 4.3 mostra os resultados obtidos.

Tabela 4.3. Resultados Validação do Modelo

Parâmetros	**Valor Machine Learning**	**Valor Estação Meteorológica**	**Variância**	**EMA do Modelo**
Temperatura [ºC]	13.227	13.215	0.230	1.043
Velocidade do Vento [km/h]	8.524	8.439	0.633	0.988
Precipitação [mm]	0.006	0.003	0.018	0.096
Evapotranspiração [mm/dia]	0.0162	0.0187	0.050	0.022
Humidade Relativa [%]	67.478	67.43	1.367	5.624
Humidade no Solo [%]	83.387	91	12.493	9.513

Como se pode observar na Tabela 4.3, para a temperatura o valor médio obtido através do ML foi de 13.227 ºC, enquanto que o valor médio da estação meteorológica foi de 13.215 ºC, o valor da variância foi de 0.230 ºC e o erro médio absoluto do modelo foi de 1.043. Para o parâmetro da velocidade do vento, o valor médio recorrendo a ML foi de 8.524 km/h, enquanto que o valor médio da estação meteorológica foi de 8.439 km/h, o valor da variância foi de 0.633 km/h e o valor de erro médio absoluto do modelo foi de 0.988 km/h. Segue-se o parâmetro da precipitação em que o valor médio recorrendo a ML, valor médio da estação meteorológica, variância e erro médio absoluto do modelo foram de 0.006 mm, 0.003 mm, 0.018 e 0.096 mm, respetivamente. A ordem dos valores para o parâmetro da

evapotranspiração foram 0.0162 mm/dia, 0.0187 mm/dia, 0.050 mm/dia e 0.022 mm/dia. Para a humidade relativa o valor médio recorrendo a ML foi 67.478%, o valor médio da estação meteorológica foi de 67.43%, o cálculo da variância foi de 1.367% e o valor do erro médio absoluto do modelo foi de 5.624%. Para concluir, o último parâmetro, para a humidade do solo o valor médio recorrendo a ML foi 83.387%, o valor médio da estação meteorológica foi de 91%, a variância foi de 12.493% e o erro médio absoluto do modelo foi de 9.513%.

Podemos então salientar que na maior parte dos casos obtivemos valores que vão ao encontro das nossas expetativas, a média das previsões está muito perto dos valores reais obtidos da estação meteorológica, sendo a diferença de 0.012 ºC na temperatura, de 0.085 km/h na velocidade do vento, de apenas 0.003 mm na precipitação, 0.0025 mm/dia na evapotranspiração, 0.048% para o parâmetro da humidade relativa e de 7.613% na humidade do solo. E os valores de erro médio absoluto do modelo foram em linha com o esperado e em conformidade uma vez que os valores da variância foram inferiores ao erro teórico no momento de treinar o modelo, como foi o caso do parâmetro da temperatura, uma vez que 0.230 ºC é menor do que 1.043 ºC, velocidade do vento também porque 0.633 km/h é menor do que 0.988 km/h, o mesmo acontece com a precipitação e com a humidade relativa. Por outro lado não aconteceu nos casos dos parâmetros da evapotranspiração e da humidade no solo, em que o valor da variância foi superior ao valor do erro médio absoluto referência do modelo, uma vez que na evapotranspiração obtivemos 0.050 mm/dia e o mesmo não é inferior a 0.022 mm/dia, valor este referente ao valor médio absoluto do modelo e na humidade do solo o valor de variância obtido como foi 12.493% então o mesmo não foi inferior ao valor do erro médio absoluto do modelo, 9.513%.

Neste capítulo foi possível escolher qual o algoritmo de machine learning e dataset que melhor se encaixa neste trabalho, bem como verificar se o mesmo cumpre os requisitos para ser utilizado, nomeadamente se o valor de erro médio absoluto não é excessivamente grande, uma vez que este valor demonstra o erro máximo que este modelo pode prever. Foi possível concluir ao longo deste capítulo que se consegue usar machine learning para prever os dados meteorológicos de forma eficiente utilizando o algoritmo floresta aleatória.

CAPíTULO 5

Modelo de Análise de Dados - Tempos e Necessidade de Rega

Como descrito na metodologia do Capítulo 3 é necessário prever os dados meteorológicos que alimentam os mecanismos de decisão através de machine learning, ao longo do Capítulo 4 foi feita toda a parte de obtenção dos dados e agora torna-se necessário implementar todo o mecanismo de decisão para os tempos de rega, desde do pedido do utilizador até à resposta final. O desenvolvimento deste capítulo é importante, uma vez que sendo este eficiente, permite escolher a melhor hora para regar num determinado intervalo e qual a duração da rega necessária, levando consequentemente a um menor desperdício de água, sendo nos dias de hoje um recurso muito importante devido à sua escassez.

Um trabalho que se destaca nesta área é o trabalho [**19**], onde o autor apresenta uma solução baseada em tecnologia IoT e ML para controlo automático de rega, utilizando nomeadamente uma rede de sensores e atuadores wireless e o autor escolheu o algoritmo XGBoost devido a este ter obtido o melhor valor de precisão, 87.73%, gerando uma poupança de 60% de água. Este trabalho será uma continuação do trabalho [**19**] focando-se mais no uso de machine learning para chegar à conclusão da necessidade de rega e nos casos de haver necessidade indicar qual a melhor data para regar e o respetivo tempo.

Há vários passos durante a análise dos dados, depois de indicados pelo utilizador, nomeadamente se a data em questão é favorável para a prática da rega recorrendo à análise dos dados anteriores, particularmente do valor da precipitação, e atuais previstos por machine learning, analisar também os dados futuros e analisar qual o melhor tempo de rega nas várias horas seguintes. Estes passos serão descritos e detalhados nas secções seguintes

5.1. Obtenção de Dados Iniciais

A obtenção dos dados é feita por uma aplicação web, em que através de uma página html são pedidos vários dados que o utilizador preenche e são passados para o script Python. A obtenção de dados pode ser feita também através de uma API, onde os dados são enviados pelo utilizador no URL para o sistema. Na Figura 5.1 pode-se observar a página referente ao tempo de rega da aplicação web.

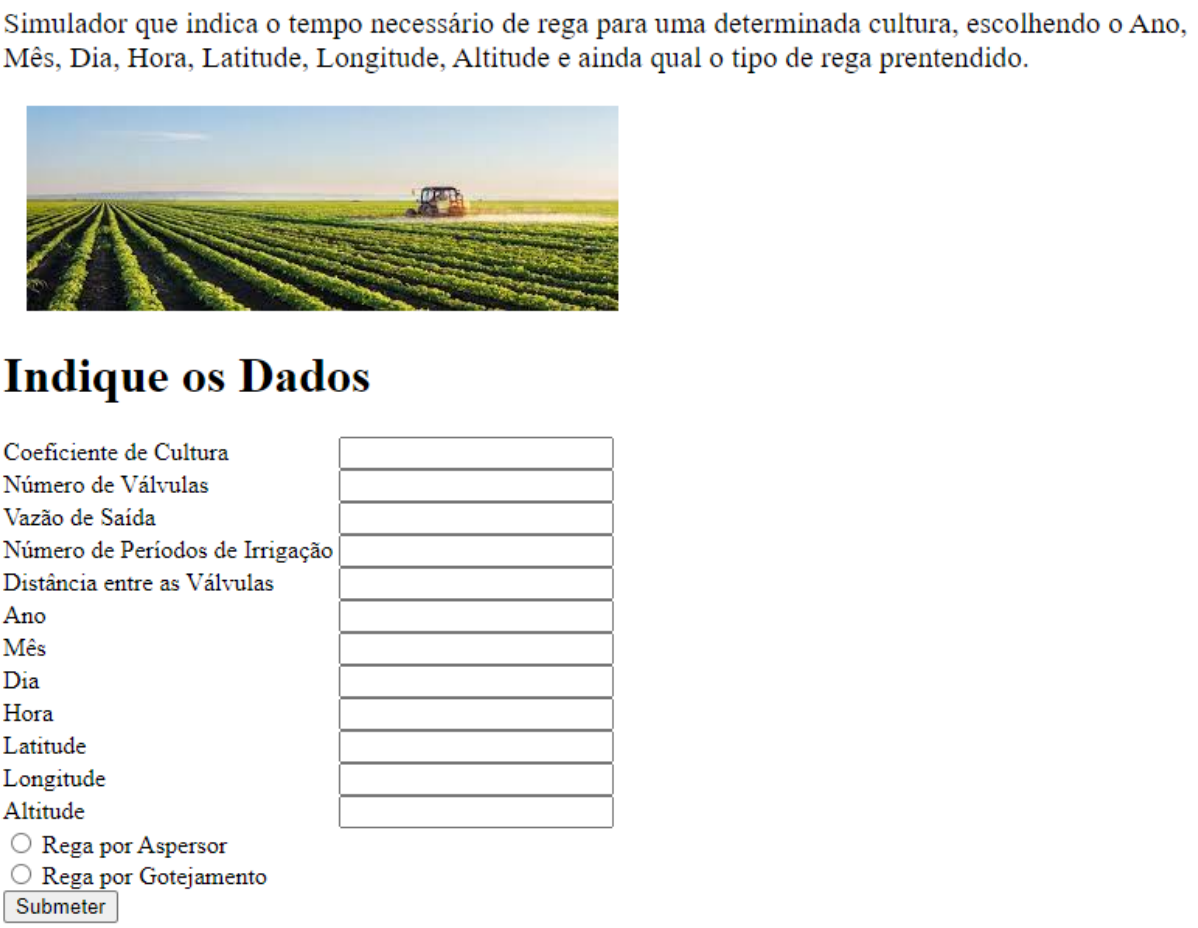

FIGURA 5.1. Tempo de Rega na Aplicação Web

Os dados obtidos são usados para previsão das condições meteorológicas e condições do solo, uma vez que os dados são usados nos modelos criados no capítulo anterior.

5.2. Análise dos Dados

Nesta secção iremos ter em conta como é calculado o tempo de rega para o dia e condições em análise, bem como a validação de se é ou não favorável regar. De maneira a analisar os dados de forma completa iremos proceder a uma comparação do tempo de rega com métodos convencionais, sem qualquer mecanismo de machine learning, com os tempos de rega obtidos através de um modelo inteligente, em que já são tidas em contas equações matemáticas de forma a obter o tempo de rega otimizado. Por fim, os dados serão analisados através de um script de python, e serão tidos em conta os melhores tempos de rega de 4 em 4 horas, uma vez que não existe necessidade de regar todas as horas, até porque pode haver momentos de chuva ou outros acontecimentos meteorológicos que indiquem que não é necessário regar. Depois de analisados os vários modelos serão

analisados os resultados dos mesmos e comparados entre si, de forma a verificar se ao usar mecanismos de ML conseguimos ou não ter melhores resultados.

5.2.1. Cálculo do Tempo de Rega

Para se calcular o tempo de rega é importante perceber que neste trabalho iremos usar dois tipos de rega, por aspersor e por gotejamento. A técnica por aspersor consiste em jatos de água lançados para o ar, acabando por se tornar em pequenas gotículas de água ao chegar ao solo, pulverizando com água os campos. Já a técnica de irrigação por gotejamento consiste, como o nome indica, na aplicação de gotas de água na planta ou cultura.

5.2.1.1. *Irrigação por Aspersor.* Foram tidas em conta uma série de fórmulas com vista a conseguir-se poupar água e com isso melhorar a sustentabilidade não só na região do campo como da água no geral. As seguintes fórmulas veem do trabalho [**30**], onde se pode ver com maior pormenor e atenção a explicação de cada parâmetro que se encontra em cada fórmula, sendo que esta tese é a continuação do trabalho referido. A Equação 5.1, permite que sejamos capazes de calcular para cada zona de irrigação o tempo necessário até ficar completamente regado tendo em contas os dados que a fórmula recebe.

$$T = \frac{A \times (K_c + ET) \times 60}{F \times N \times 1000} / P \tag{5.1}$$

Na Equação 5.1 é tido em conta o tipo de válvulas que são usadas no processo da irrigação e a distância entre as mesmas, são usados parâmetros tais como a humidade do solo, a temperatura do ar e a humidade relativa do ar, estando estes subentendidos na fórmula, K_c é o coeficiente da cultura, sendo um valor definido para cada tipo de cultura. A quantidade de válvulas é definida por N, ET é a evapotranspiração, usada na unidade de medida em mm/dia e a sua fórmula pode ser observada na Equação 5.3, já o parâmetro P serve para ser possível dividir pelo número de períodos de irrigação que é desejável ter, já F refere-se à vazão da água e encontra-se na unidade de medida m^3/h. Por fim A, pode ser observado como se calcula na Equação 5.2 e refere-se à área em m^2 do campo.

$$A = [(0.5 \times N) - 1] \times D^2 \tag{5.2}$$

Na Equação 5.2, N é a quantidade de válvulas no campo e D é a distância entre as mesmas.

Na Equação 5.3, pode-se observar como se calculou a evapotranspiração, tendo sido usada a fórmula de Hargreaves simplificada [**31**].

$$ET = 0.0023 \times (T_{med} + 17.78) \times Ro \times (T_{max} - T_{min})^{0.5} \tag{5.3}$$

Na Equação 5.3, T_{med} refere-se à temperatura média, sendo que Ro significa a radiação solar extra-terrestre incidente, geralmente a unidade é expressa em mm/dia. T_{max} e T_{min} dizem respeito à temperatura máxima e mínima, respetivamente.

$$TI_x = \frac{T \times (100 - I_x)}{100} \tag{5.4}$$

Na Equação 5.4, pode ser observado como se calcula o tempo genérico até chegar a 100% da variável x, sendo esta equação usada para calcular o TI_{solo} e também o TI_{hum}. I_x é o valor lido pelo sensor.

Por fim, otimizou-se a Equação 5.1 para refletir a chuva que esteve nas últimas 24 horas, desta forma a fórmula final é a Equação 5.5.

$$T = \frac{A \times (K_c + ET - TI_{24chuva}) \times 60}{F \times N \times 1000} / P \tag{5.5}$$

A Equação 5.6 representa um valor de tempo otimizado tendo em conta valores como TI_{solo} e TI_{hum} e o seu cálculo pode ser observado na Equação 5.4. O parâmetro $TI_{24chuva}$ indica a quantidade de precipitação que ocorreu nas passadas 24h, quanto maior este valor, menor o tempo de rega.

$$Topt = TI_{solo} \times 0.7 + TI_{hum} \times 0.3 + T \times 0.1 \tag{5.6}$$

Por fim, na Equação 5.6 podemos ver como é calculado o tempo de rega otimizado, indicando o tempo necessário que é preciso regar. Esta equação necessita dos valores que já vimos nas equações imediatamente acima, como por exemplo do valor T que é o resultado da Equação 5.5, TI_{solo} e TI_{hum} que provêm da Equação geral 5.4.

5.2.1.2. *Irrigação por Gotejamento.* Tal como na irrigação por aspersor, existem também uma série de fórmulas para a irrigação utilizando esta técnica, sendo elas bastantes semelhantes e algumas são exatamente iguais, a Equação 5.7 indica o tempo em minutos de irrigação.

$$T = \frac{A \times (K_c + ET) \times 60}{\frac{N}{D} \times F} / P \tag{5.7}$$

Na Equação 5.7 os parâmetros que a fórmula tem são todos parâmetros que a Equação 5.1 já tinha.

De seguida é possível observar a Equação 5.8, que demonstra como se pode calcular a área da rega, tendo em conta que a irrigação é feita de outra maneira e consequentemente a fórmula para o seu cálculo também é diferente.

$$A = N \times D \tag{5.8}$$

Na Equação 5.8, N é a quantidade de válvulas no campo e D é a distância entre as mesmas.

As fórmulas relativamente à evapotranspiração e ao TI são exatamente iguais ao tipo de rega por aspersor que foi apresentado imediatamente acima, nas Equações 5.3 e 5.4, respetivamente.

$$T = \frac{A \times (K_c + ET - TI_{24chuva}) \times 60}{\frac{N}{D} \times F} / P \tag{5.9}$$

A Equação 5.10 representa um valor de tempo otimizado tendo em conta valores como TI_{solo} e TI_{hum} e o seu cálculo pode ser observado na Equação 5.6. O parâmetro $TI_{24chuva}$ diz respeito à precipitação que ocorreu ao longo das 24h passadas, sendo que se este valor for grande não haverá tanta ou nenhuma necessidade de rega.

$$Topt = TI_{solo} \times 0.7 + TI_{hum} \times 0.3 + T \times 0.1 \tag{5.10}$$

Na Equação 5.10 obtém-se o tempo em minutos que é necessário regar o campo utilizando para isso o valor final resultante da Equação 5.9.

5.2.1.3. *Condições Favoráveis à Rega.* Teve-se em consideração uma série de regras à rega, nomeadamente se alguns dados meteorológicos estão dentro de vários parâmetros, de forma a não regar caso não seja favorável ou se não for necessário. Assim, só se considerou viável a rega caso a velocidade do vento fosse inferior a 36 km/h ou velocidade do vento inferior a 25 km/h e direção do vento entre 180 º e 359 º inclusive, ao mesmo tempo a precipitação ser inferior a 1 mm e a temperatura ser inferior a 30 ºC, caso todas estas regras fossem cumpridas então estamos perante uma data favorável para a rega. Caso estejamos perante um caso que é favorável regar, mas o tempo calculado for inferior a zero, então nesses casos será arredondado para zero uma vez que não há necessidade de regar. Em todos os outros cenários, a rega foi considerada desfavorável.

5.2.2. Script de Análise de Dados

Depois de apresentadas as fórmulas matemáticas, procedeu-se à incorporação das mesmas num script Python. Este apresenta vários mecanismos de inteligência de forma a conseguir calcular o melhor tempo de rega caso haja necessidade para tal.

Assim que um pedido chega à aplicação web, o script começa por fazer algumas validações utilizando os dados indicados pelo utilizador, por exemplo se o tipo de rega não for nenhuma das esperadas então o utilizador recebe uma mensagem que tem de preencher corretamente esse campo. Caso haja campos vazios, o mesmo acontece.

Se tudo estiver correto, o script começa por prever os dados meteorológicos da data prevista e das 4 horas posteriores, uma vez existem condições que indicam se a rega é favorável, as datas que sejam favoráveis ficam guardadas para prosseguirem ao longo do script, enquanto que as desfavoráveis são apagadas.

De forma a obtermos um método, utilizando machine learning, mais eficiente, usou-se a melhor hora para regar durante um intervalo de 4 horas, isto é, analisa-se se ao longo destas 4 horas de forma a verificar se irá existir algum acontecimento meteorológico que indique que não é necessário regar até essa data, sendo feito este raciocínio todas as horas. Desta forma parte-se do princípio que não é necessário regar todas as horas e com esta medida tenta-se regar na hora mais vantajosa e que seja realmente necessário regar, sendo a rega feita durante menos tempo e desta forma levar a uma menor quantidade de água desperdiçada.

De seguida, utilizam-se as várias fórmulas respetivas ao tipo de rega. Para obter o valor necessário para a chuva nas 24 horas anteriores, são somados os valores previstos para a chuva nas 24 horas anteriores. À medida que vão sendo calculados cada um dos tempos otimizados, esses são verificados com os anteriores de forma a ser analisada qual a melhor data para indicar ao utilizador, caso haja um tempo otimizado igual a 0, então significa que naquela hora poderá chover ou existir alguma mudança climatérica que não justifica regar até essa data inclusive, e parte-se para as datas posteriores.

No final de analisar todas as datas, o script obtém uma data favorável e em que seja precisa menos água para regar ou caso não haja nenhuma data favorável, o script termina com a conclusão que se deve esperar mais do que 4 horas para se proceder à rega.

5.3. Apresentação dos resultados

Os resultados obtidos pelo script são apresentados ao utilizador através de uma resposta por parte da aplicação web, caso o script não tenha encontrado uma data favorável

para o utilizador regar, então é apresentada a resposta: "Devido às condições que estarão nas próximas 4 horas não é necessário regar, deverá aguardar para regar". Caso contrário, a aplicação apresenta a resposta: "Deverá regar na seguinte data (Ano,Mes,Dia,Hora): "e indica de seguida a data constituída pelo ano, mês, dia e hora que melhor se adequa para a rega.

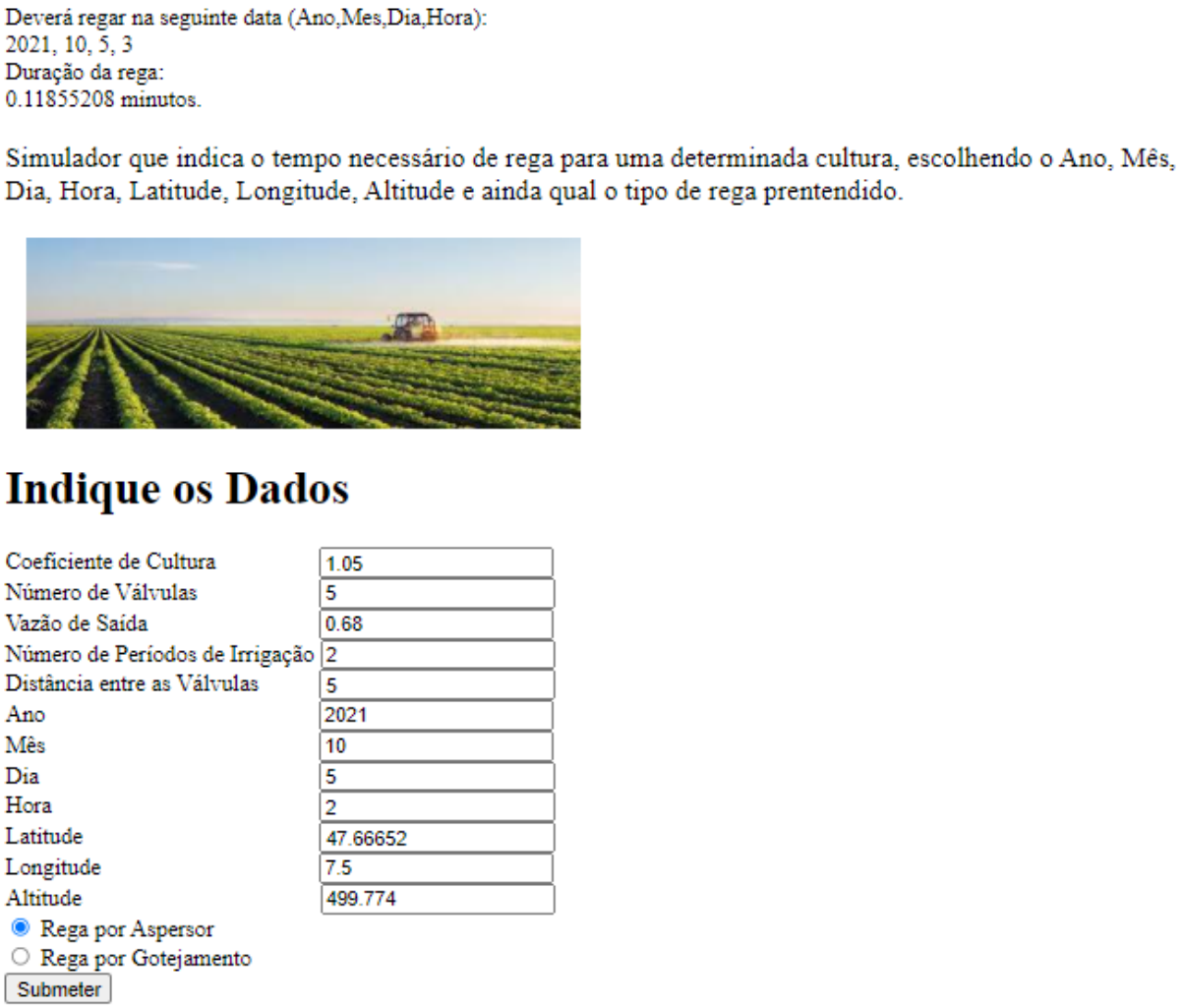

FIGURA 5.2. Resultado do Pedido do Utilizador na Aplicação Web

Como se pode verificar na Figura 5.2 os resultados são apresentados na mesma página do pedido inicial feito pelo utilizador.

No caso de o pedido ter sido feito através da API, a resposta é apresentada num JSON, contendo as mesmas informações, tal como se pode ver na Figura 5.3.

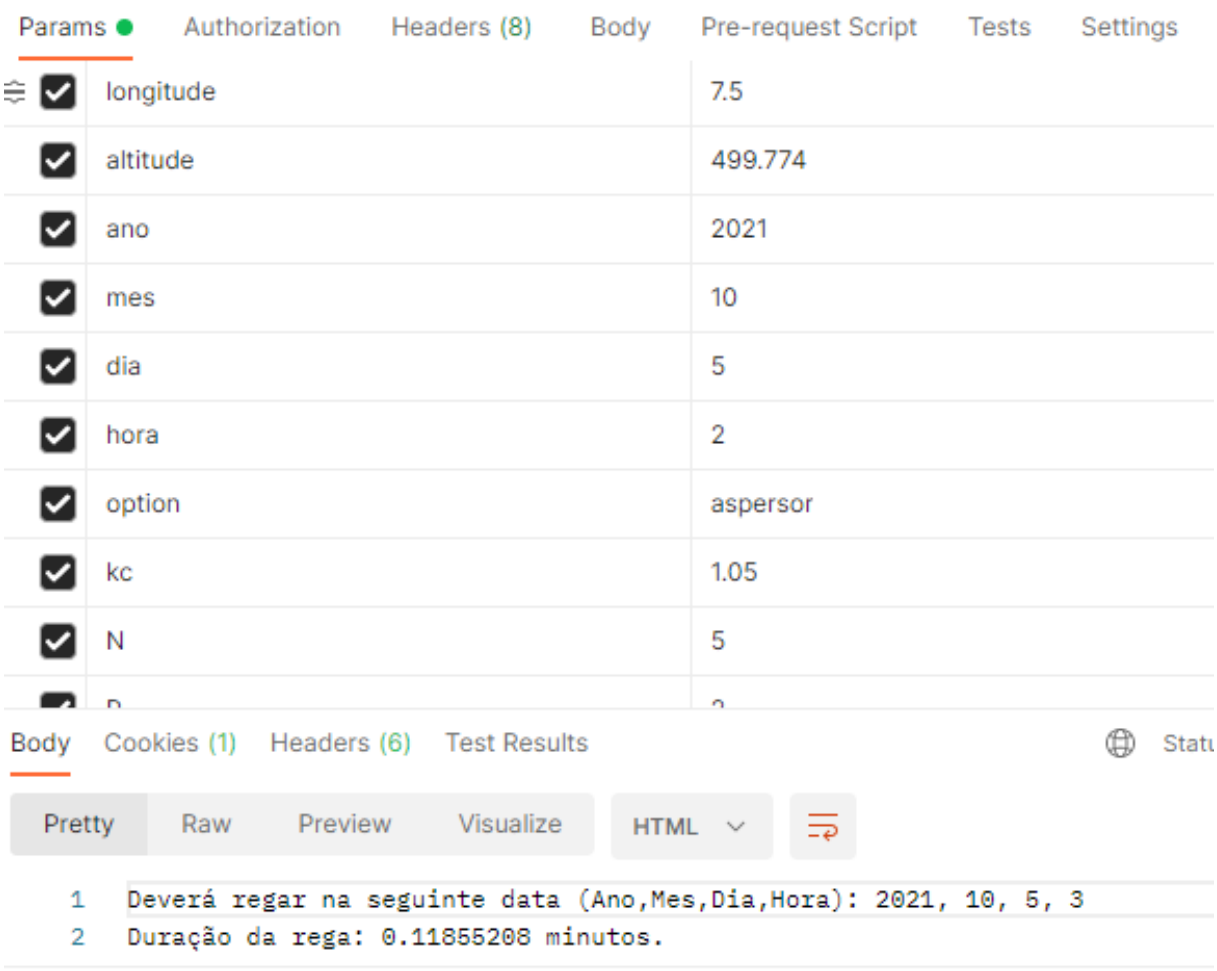

FIGURA 5.3. Resultado do Pedido do Utilizador na API

CAPíTULO 6

Validação do Sistema

Após desenvolvimento do sistema proposto, procedeu-se à validação do mesmo de modo a compreender se, não só o mesmo é capaz de cumprir com os objetivos propostos, nomeadamente prever as melhores horas e tempos de rega baseado em mecanismos de machine learning, mas também perceber como se comporta quando comparado com outros cenários de rega, sejam cenários convencionais ou cenários inteligentes. Esta validação permite também compreender se o mesmo consegue ou não ser mais eficiente e sustentável.

Para proceder a esta validação foram realizados um conjunto de testes que, tal como os seus resultados e a discussão dos mesmos, serão apresentados nas secções seguintes.

6.1. Cenários de Teste

De forma a validar qual a melhor maneira para prever o tempo de rega foram pensados 3 cenários de testes:

- **Cenário Convencional:** Neste cenário foi considerado um método convencional de rega, em que são usadas as informações provenientes das estações meteorológicas, não havendo qualquer mecanismo de inteligência, juntamente com as Equações 5.1 e 5.7, sendo a primeira fórmula usada no contexto da rega por aspersor e a segunda fórmula para o tipo de rega gota a gota, para calcular o tempo de rega para o dia em questão. Este é o cenário que podemos encontrar em muitos campos agrícolas, maioritariamente naqueles em que os agricultores apresentam menos capacidades tecnológicas e mais experiência tradicional, tendo trabalhado assim toda a sua vida. Visto haver experiência por parte dos agricultores, assumiu-se que nas horas que a humidade do solo seja 90% ou maior não é necessário regar.
- **Cenário Inteligente:** Neste cenário foi considerada o sistema de rega inteligente apresentado no trabalho [**19**], onde são usados uma mistura de dados das estações meteorológicas e de sensores colocados no campo agrícola. Para o cálculo dos tempos de rega são usadas as condições apresentadas de forma a averiguar se a rega é favorável naquela data, juntamente com as fórmulas otimizadas para o

cálculo do tempo de rega, como já apresentado no Capítulo 4, as fórmulas para o tipo de rega por aspersor e para o tipo de rega gota a gota são as Equações 5.6 e 5.10, respetivamente. No trabalho anteriormente referido, este cenário obteve melhorias face ao método convencional de 60% em poupança de água. Este cenário usa mecanismos de machine learning, mas apenas na decisão da rega, fazendo um uso parcial das capacidades de machine learning e por isso indicado como cenário inteligente.

- **Cenário com Machine Learning:** Neste último cenário foi considerada o sistema desenvolvido nesta tese, completamente baseado em machine learning, e onde os valores das condições meteorológicas e do solo são previstos através de modelos e depois aplicados nas fórmulas otimizadas, como apresentado no Capítulo 5. O objetivo deste cenário é demonstrar se também é possível utilizar técnicas de machine learning para calcular o tempo de rega baseado em dados obtidos em períodos de rega anteriores, de modo a substituir algumas das equações apresentadas anteriormente e os sensores necessários no cenário inteligente. Tal como identificado, este cenário foi apresentado como sendo o cenário que recorre a mecanismos de machine learning, uma vez que efetivamente usa machine learning para substituir toda a parte dos sensores do cenário inteligente.

Dada a impossibilidade de testar estes cenários num ambiente real, foi simulado um campo agrícola e usados os dados obtidos tanto no dataset horário, indicado no Capítulo 4, bem como os dados obtidos na implementação do trabalho associado ao cenário inteligente [**19**]. Assim sendo, foi simulado um ano inteiro de rega, neste caso foi usado o ano de 2020 em todos os cenários, pois este ano não foi usado para treinar os modelos de machine learning. Para cada um dos cenários foi calculado, a cada hora, a necessidade de rega baseado no modelo utilizado em cada cenário. De modo a analisar o impacto do método utilizado foram feitas as médias por hora do tempo necessário para irrigar um determinado campo simulado, ao longo de todo o ano de 2020.

6.2. Resultados

A Tabela 6.1 apresenta os resultados obtidos das médias do tempo que é necessário irrigar por hora para o ano de 2020 em cada um dos cenários testados.

Tabela 6.1. Média do Tempo de Irrigação por hora

Técnica de Irrigação	**Média do Tempo de Irrigação por hora [minutos]**		
	Cenário Convencional	**Cenário Inteligente**	**Cenário ML**
Aspersor	0.111	0.039	0.031
Gotejamento	0.698	0.316	0.194

É possível observar que para a técnica por aspersor o valor da média obtido para o cenário convencional foi de 0.111 minutos, no cenário inteligente foi de 0.039 minutos e para o cenário que usa machine learning obteve-se 0.031 minutos. Portanto houve uma redução de 64.86% do tempo de rega utilizando o cenário inteligente ao invés do cenário convencional, utilizando o cenário com machine learning obteve-se uma melhoria de 20.51% em relação ao cenário inteligente. Já utilizando o cenário que recorre a machine learning ao invés do cenário convencional obtivemos uma melhoria de 72.07% de água.

Para a técnica por gotejamento obteve-se 0.698 minutos para o cenário convencional, 0.316 minutos para o cenário inteligente e para o cenário que recorre a ML obteve-se o valor médio de 0.194 minutos. Obteve-se uma redução do tempo de rega de 54.73% do cenário convencional para o cenário inteligente e utilizando o cenário com machine learning ao invés de usar o cenário inteligente, a redução foi de 38.61%. Utilizando o cenário com machine learning obtivemos uma redução de água de 72.21% ao invés de usarmos o cenário convencional.

Em comparação com o trabalho [**19**], podemos observar que o autor obteve valor de redução de água na ordem dos 60% e este trabalho obteve uma redução de 64.86%, sendo o mesmo ligeiramente melhor.

É possível de se perceber, através da análise da Tabela 6.1 que independentemente da técnica de irrigação o cenário que poupa mais água é o cenário que recorre a machine learning, seguido pelo cenário inteligente, com uma melhoria de 20%, e por fim, como seria de esperar, o cenário convencional, com uma eficiência de 60% inferior aos cenários inteligentes.

6.3. Custos

Apesar da poupança em termos de água, é importante considerar que a utilização de cenários inteligentes acarreta custos associados que não existem nos cenários convencionais, nomeadamente na instalação de sensores e na troca de mensagens com servidores. E um sistema não pode ser considerado eficiente, se o custo de manutenção e instalação for superior às poupanças que apresenta.

Os custos associados a cada cenário são diferentes, uma vez que cada cenário tem diferentes componentes associados e o custo associado aos mesmos tem um peso decisivo na hora de decidir qual o cenário a optar.

No cenário convencional não serão considerados custos, uma vez que as decisões partem da sabedoria do agricultor, não havendo quaisquer aparelhos eletrónicos e nesse sentido este cenário não será considerado para esta secção.

De forma a serem implementados os cenários, torna-se importante analisar e comparar os custos associados a cada um. Numa ótica de equipamento, no cenário inteligente existirão equipamentos tais como, sensores, controladores, a necessidade de uma estação meteorológica e de um servidor.

A implementação da solução apresentada nesta dissertação, em termos de equipamento, é mais simples que o cenário inteligente, uma vez que não necessita de sensores nem de estação meteorológica.

Tabela 6.2. Custos Associados a Cada Cenário

<table>
<tr><th></th><th>Equipamento</th><th>Cenário Inteligente</th><th>Cenário com ML</th></tr>
<tr><td rowspan="4">Hardware</td><td>Servidor [€/mês]</td><td>30</td><td>70</td></tr>
<tr><td>Estação Meteorológica [€/hectare]</td><td rowspan="3">3000</td><td>0</td></tr>
<tr><td>Controladores [€/hectare]</td><td>500</td></tr>
<tr><td>Sensores [€/hectare]</td><td>0</td></tr>
<tr><td>Software</td><td>Software [€]</td><td>0</td><td>125</td></tr>
</table>

Na Tabela 6.2 pode-se observar os custos associados a cada cenário, como já vimos anteriormente, o cenário que recorre a machine learning não necessita de sensores nem de estação meteorológica e por isso os valores destes equipamentos é de 0€. O valor do servidor é de 30€ por mês no cenário inteligente e de 70€ por mês no cenário com machine learning uma vez que no cenário inteligente não é necessário um servidor muito complexo, visto não existir uma grande necessidade de poder computacional, já para o cenário com machine learning será necessário um servidor mais complexo e com maior poder computacional, para implementar os modelos necessários. Outra diferença entre os dois cenários é no custo inicial para implementar os modelos de análise e os scripts de software, para o cenário inteligente não é necessário qualquer software adicional, enquanto que para o cenário que usa machine learning é necessário adquirir os dataset necessários para treinar os vários modelos, de modo a substituir os sensores.

Na Tabela 6.3 pode-se observar os custos ligados à parte da manutenção de cada um dos cenários.

Tabela 6.3. Custos de Manutenção

Cenário	**Manutenção [€]**
Inteligente	**300**
Utilizando ML	**50**

Para o cenário inteligente estimou-se um valor de manutenção anual de 300€, enquanto que para o cenário que recorre a machine learning estimou-se um valor de 50€ de manutenção anual. Estes valores, no caso de ser o cenário inteligente, será o valor da mudança de pilhas, a verificação do estado dos sensores e a sua substituição caso seja necessário, no caso do cenário com ML estimou-se um valor mais baixo, uma vez que não existem sensores, portanto, basta garantir que o sistema está funcional, sendo por isso mais barato.

Estes valores quando comparados ao fim de 1 ano, podem ser observados na Tabela 6.4.

Tabela 6.4. Custos Anuais

Cenário	**Custos Totais no 1º ano [€]**	**Custos por Ano Posterioremente [€]**
Inteligente	3660	660
Utilizando ML	1515	890

Nos custos anuais para os 2 cenários está incorporado o valor mensal do servidor, multiplicado pelos 12 meses e somado o valor de manutenção respetivo que se viu na Tabela 6.3. Depois de somados os valores, obteve-se para o cenário inteligente o valor de 660€ anuais e para o cenário que recorre a machine learning obteve-se 890€.

Como se pode observar, o custo inicial para o cenário inteligente é de 3660€, enquanto que para o cenário que recorre a machine learning é de 1515€. Neste valor inicial são inseridos os valores iniciais que são mostrados na Tabela 6.2, somados com os custos de manutenção indicados na Tabela 6.3, havendo uma diferença de 2145€ no total do primeiro ano. Nos anos seguintes este valor passa a ser mais similar, sendo no caso do cenário inteligente de 660€ e no cenário que recorre a machine learning é de 890€, a diferença é de 230€.

Com estes dados podemos concluir que o investimento inicial é mais económico no caso do cenário que recorre a machine learning e apesar de apresentar os valores anuais posteriores mais elevados que o cenário inteligente, até estes ficarem sensivelmente parecidos monetariamente teriam de passar 9,33 anos e após este período o cenário mais barato passaria a ser o cenário inteligente. No entanto como visto na secção dos resultados deste capítulo, o cenário que poupa mais água é o cenário que recorre a machine learning e por isso será o que gasta menos água ao longo de todos os anos.

De modo a perceber se o valor da instalação e manutenção de um sistema inteligente para a rega compensa financeiramente face à quantidade de água poupada, i.e., anualmente o sistema não fica mais caro que o valor de água poupado, foi necessárioo perceber qual a redução no consumo anual.

A seguinte equação representa o consumo dos litros por hora, Equação 6.1.

$$C = \frac{(F \times N \times 1000) \times T}{60} \tag{6.1}$$

Na Equação 6.1 podemos observar que o consumo depende de F que é a vazão da água, depende de N que representa o número de válvulas e também do tempo que é necessário regar. Na seguinte tabela pode-se observar os valores do consumo de água em litros por hora para a técnica de irrigação por aspersor.

Tabela 6.5. Consumo de Água - Aspersor

Cenário	**Consumo [L/h]**
Convencional	6.29
Inteligente	2.21
com ML	1.76

Sendo que o preço do litro da água no alentejo será na ordem dos 0,3€ [**32**], então somos capazes de calcular o valor gasto por hora em cada um dos cenários como apresentado na Tabela 6.5. No caso do cenário convencional será pago por hora o valor de 1,89€, no caso do cenário inteligente será pago 0,66€, enquanto que no cenário que recorre a machine learning será de 0,53€. Sendo que para calcularmos por anos, basta multiplicar por 24, sendo as horas de um dia e por 365 para ficarmos com o custo anual, os valores serão 16554,4€, 5781,6€ e 4642,8€, para os cenários convencional, inteligente e que recorre a machine learning, respetivamente.

A fórmula para a técnica por gotejamento usada no cálculo do consumo da água é ligeiramente diferente da Equação 6.1, uma vez que a vazão, F, já se encontra nas

unidades L/h e por esse motivo já não é necessário passar de m^3/h para L/h. Foi utilizada a Equação 6.2.

$$C = \frac{(F \times N) \times T}{60} \tag{6.2}$$

Tabela 6.6. Consumo de Água - Gotejamento

Cenário	**Consumo [L/h]**
Convencional	3.14
Inteligente	1.42
com ML	0.87

Com a técnica de irrigação gota a gota, os dados podem ser observados na Tabela 6.6. No caso do cenário convencional será pago por hora o valor de 0,94€, no caso do cenário inteligente será pago 0,43€, enquanto que no cenário que recorre a machine learning será de 0,26€. Os custos anuais, serão 8234,4€, 3766,8€ e 2277,6€, para os cenários convencional, inteligente e que recorre a machine learning, respetivamente.

Depois de calculados os valores de água gastos anualmente em cada um dos cenários, podemos constatar que apesar de haver gastos no cenário inteligente e no cenário que recorre a machine learning, mesmo assim o que estes dois cenários permitem em termos de poupança de água é muito mais vantajoso e calculando os valores gastos da água e do consumo em cada cenário, chegamos à conclusão que mesmo antes do término do primeiro ano ambos os cenários inteligentes são mais vantajosos economicamente.

6.4. Discussão

Como foi possível de analisar, quando comparados os 3 cenários usados nesta validação, o que sobressai pela negativa é o cenário convencional, sendo o que apresentou piores resultados. O cenário inteligente apresenta melhores resultados, como analisado anteriormente, em que foi possível obter valores na ordem dos 60% em termos de redução do tempo de rega e o cenário que sobressai pela positiva é o cenário desenvolvido neste trabalho, o que demonstra que apesar de poder haver algum erro associado à previsão dos vários dados meteorológicos, este é o cenário que consegue usar menos água. É então possível reduzir o desperdício, em comparação com os outros cenários, sendo obtida uma redução de água na casa dos 72% em relação ao cenário convencional. Também foi possível obter melhores resultados no cenário que recorre machine learning, uma vez que se usou o raciocínio de analisar e calcular qual a melhor hora, sendo o raciocínio feito de 4 em 4 horas, desta forma conseguiu-se evitar a irrigação todas as horas como nos outros cenários.

Em relação aos custos podemos observar que o melhor cenário, mais uma vez, é o cenário desenvolvido ao longo deste trabalho, uma vez que apresenta menores custos em relação ao do cenário inteligente e o que apresenta menor consumo de água e, portanto, independentemente da técnica de irrigação usada, o melhor cenário a optar é o cenário que recorre a machine learning, desenvolvido neste trabalho, sendo possível popar-se 10396,6€ ou 4441,8€ logo após o primeiro ano, ao invés de se usar o cenário convencional e recorrendo à técnica por aspersor ou por gotejamento, respetivamente.

CAPíTULO 7

Conclusões e Trabalho Futuro

7.1. Principais Conclusões

Nesta dissertação foi apresentado um sistema capaz de decidir qual a melhor hora para regar num determinado dia. Este trabalho tem como objetivo a maior poupança de água possível e para isso a solução foi implementada através de um algoritmo de machine learning capaz de fazer a previsão dos dados meteorológicos dessas mesmas horas e escolher a hora mais vantajosa tendo em conta a que necessita da menor quantidade de água para regar o campo.

Primeiramente procedeu-se ao estudo do conceito de machine learning, nomeadamente ligado à agricultura, dos tipos de aprendizagem e dos algoritmos. Também se procedeu ao estudo de alguns trabalhos na área. De seguida foi analisado e feita a metodologia do que seria feito posteriormente.

Ao longo do trabalho para se proceder à previsão dos dados meteorológicos percebeu-se que podiam ser usados algoritmos de machine learning. Tendo isso em conta foram testados quatro algoritmos de aprendizagem supervisionada utilizando um dataset diário, o algoritmo floresta aleatória, redes neuronais, árvore de decisão e regressão linear, sendo que os algoritmos que apresentaram melhores resultados de erro médio absoluto foram os algoritmos floresta aleatória e árvore de decisão. Foi feito um segundo estudo para perceber a diferença entre a utilização de dados diários e dados horários, sendo que neste teste, procedeu-se apenas com os melhores algoritmos do teste anterior. Depois de analisados os resultados utilizando o dataset horário, concluiu-se que o melhor algoritmo para prever os dados meteorológicos foi a floresta aleatória e que o dataset mais indicado seria o dataset horário. Procedeu-se ainda à validação do modelo de forma a verificar a aceitabilidade do mesmo, que consistiu em verificar se os valores de erro estariam dentro do intervalo permitido, em que verificou-se que para a temperatura, a variância obtida foi de 0.230 ºC, enquanto que o valor do erro médio absoluto do modelo foi de 1.043 ºC, para a velocidade do vento, os valores foram 0.633 km/h e 0.988 km/h, variância e erro médio absoluto respetivamente. Para a precipitação foram obtidos os valores 0.018 mm e 0.096 mm, variância e erro médio absoluto, respetivamente. Para a evapotranspiração,

o valor da variância obtido foi de 0.05 mm/dia e de 0.022 mm/dia para o erro médio absoluto. Para a humidade relativa foi obtido o valor de 1.367% de variância e de 5.624% de erro médio absoluto e por fim, para a humidade no solo foi obtido o valor de variância de 12.493% e o valor 9.513% de erro médio absoluto do modelo. Com estes resultados de validação foi possível observar que em 7 parâmetros, apenas 2 saiem ligeiramente do valor esperado máximo em relação ao valor obtido, sendo eles a evapotranspiração e a humidade no solo.

De seguida realizou-se a implementação do sistema, com base em dados indicados pelo utilizador, conseguiu-se através de um script calcular se havia necessidade de rega e em caso positivo, indicar a melhor hora e a duração da rega mais sustentável. Para isto, foi necessário analisar a necessidade de rega utilizando algumas regras que indicam se será vantajoso ou não a rega, tendo em conta os dados previstos. Para calcular o tempo de rega usaram-se fórmulas matemáticas e no fim ainda foi necessário analisar qual das horas seria mais conveniente ou se devido aos dados meteorológicos futuros se nem existiria necessidade de rega. Percebeu-se durante o desenvolvimento do mesmo que prever durante 24 horas a necessidade da rega seria um intervalo demasiado longo e por esse motivo usou-se um intervalo de 4 horas, visto ir mais ao encontro das necessidades reais dos campos agrícolas. Procedeu-se aos cálculos para dois tipos de rega, por aspersor e por gota a gota.

Realizou-se a comparação de 3 cenários diferentes de rega, de forma a validar se o sistema desenvolvido seria mais ou menos vantajoso em relação a outros cenários mais comuns, inclusive comparando os custos, tanto em relação à água gasta em cada cenário como também ao preço dos equipamentos e manutenção dos mesmos. Foi possível concluir que o método que permite uma maior eficiência foi o método desenvolvido nesta dissertação, que recorre a machine learning. Uma vez que recorrendo a este método é possível uma poupança ao fim do primeiro ano de 10396,6€ se for usada a técnica de irrigação por aspersor ou de 4441,8€ caso seja por gotejamento, isto utilizando o cenário que recorre a machine learning ao invés do método convencional. Ao longo dos anos seguintes a poupança aumenta, uma vez que os custos anuais posteriores já não incluem os custos iniciais necessários a ambos os cenários inteligentes, sendo apenas necessário fazer a manutenção.

Em relação ao trabalho [**19**], foi possível observar que com o uso de machine learning conseguiu-se uma redução de água de 72%, que foi mais 12% do que os autores do trabalho

conseguiram utilizando um método inteligente sem o mesmo uso de machine learning como apresentado neste trabalho.

Em suma o trabalho cumpriu com os objetivos gerais propostos no início do mesmo, uma vez que no final conseguiu-se criar um sistema capaz, de forma autónoma, de prever dados meteorológicos utilizando um algoritmo de machine learning e capaz de indicar ao utilizador, consoante o local e o dia, a necessidade de rega naquela hora, ou se seria mais favorável regar depois, indicando o tempo essencial para a rega ser concluída. O sistema foi capaz de gerar poupança de água e apesar dos custos inerentes que o sistema apresenta, foi possível gerar um sistema mais económico que o cenário convencional apresenta antes do fim do primeiro ano.

7.2. Trabalho Futuro

Para trabalho futuro existem algumas melhorias que podem ser implementadas:

- **Cenário Real:** Testar num cenário real o sistema implementado de forma a poder verificar-se que de facto o sistema comporta-se de forma eficiente.
- **Aplicação Móvel:** Desenvolver uma aplicação para sistemas Android e IOS possibilitaria a facilidade de aceder à aplicação mais rapidamente, não havendo necessidade de colocar sempre o link no motor de busca para a página web.
- **Aumentar Funções do Sistema:** Pode-se desenvolver novas funcionalidades do sistema, tais como, prever os dados meteorológicos num tempo mais duradouro e desta forma fazer a previsão mais do que as 4 horas atuais.

Referências

[1] H. N. Saha, A. Mandal, and A. Sinha, "Recent trends in the internet of things," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, 2017, pp. 1–4.

[2] K. Asthon, "That ' internet of things ' thing," 2010, [Online] Available: http://www.rfidjournal.com/article/print/4986, (visited 15/02/2021).

[3] G. Simões, C. Dionísio, A. Glória, P. Sebastião, and N. Souto, "Smart system for monitoring and control of swimming pools," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019, pp. 829–832.

[4] "Number of connected devices worldwide 2030 — statista." 2020, [Online] Available: https://www.statista.com/statistics/802690/worldwide-connected-devices-by-access-technology/, (visited 02/01/2021).

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[6] J. A. Stankovic, "Research directions for the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, 2014.

[7] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N. Lekbangpong, A. Wanichsombat, and P. Nillaor, "Iot and agriculture data analysis for smart farm," *Computers and electronics in agriculture*, vol. 156, pp. 467–474, 2019.

[8] K. Lakhwani, H. Gianey, N. Agarwal, and S. Gupta, *Development of IoT for Smart Agriculture a Review: Proceedings of ICETEAS 2018*, 01 2019, pp. 425–432.

[9] H. S and R. Ramathmika, "Sentiment analysis of yelp reviews by machine learning," 05 2019, pp. 700–704.

[10] R. R. Reddy, C. Mamatha, and R. G. Reddy, "A review on machine learning trends, application and challenges in internet of things," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 2389–2397.

[11] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1310–1315.

[12] D. Sharma and N. Kumar, "A review on machine learning algorithms, tasks and applications," vol. 6, pp. 2278–1323, 10 2017.

[13] J. M. d. J. A. COELHO, "Machine learning for precise water leaks detection," Master's thesis, Lisboa: Iscte, 2020.

[14] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, p. 2674, 08 2018.

[15] M. Shakoor, K. Rahman, S. Rayta, and A. Chakrabarty, "Agricultural production output prediction using supervised machine learning techniques," 07 2017, pp. 182–187.

[16] R. Amami, D. Ayed, and N. Ellouze, "Practical selection of svm supervised parameters with different feature representations for vowel recognition," vol. 7, 07 2015.

[17] M. Goyal, C. Ojha, and D. Burn, *Machine Learning Algorithms and Their Application in Water Resources Management*, 10 2017, pp. 165–178.

[18] S. Mishra, D. Mishra, and G. Santra, "Applications of machine learning techniques in agricultural crop production: A review paper," *Indian Journal of Science and Technology*, vol. 9, 10 2016.

[19] J. M. B. CARDOSO, "Smartfarm: improve sustainability using wireless sensor networks," Master's thesis, Lisboa: Iscte, 2020.

[20] S. Georganos, T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, and E. Wolff, "Very high resolution object-based land use-land cover urban classification using extreme gradient boosting," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, pp. 1–5, 02 2018.

[21] X. Yu, "Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china," *Energy Conversion and Management*, vol. 164, 03 2018.

[22] S. Ding, Z. Zhibin, and X. Zhang, "An overview on semi-supervised support vector machine," *Neural Computing and Applications*, vol. 28, 05 2017.

[23] B. Jang, M. Kim, G. Harerimana, and J. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. PP, pp. 1–1, 09 2019.

[24] T. Anjali, K. Chandini, K. Anoop, and V. L. Lajish, "Temperature prediction using machine learning approaches," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1, 2019, pp. 1264–1268.

[25] T. Saba, A. Rehman, and J. AlGhamdi, "Weather forecasting based on hybrid neural model," *Applied Water Science*, 02 2017.

[26] M. Holmstrom and D. Liu, "Machine learning applied to weather forecasting," 2016.

[27] S. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic, and S. Alessandrini, "Machine learning for applied weather prediction," 10 2018, pp. 276–277.

[28] "Meteoblue," [Online] Available: https://content.meteoblue.com/pt/formas-de-acesso/api-meteorologica-de-meteoblue, (visited 10/01/2021).

[29] "Scikit-learn, joblib," [Online] Available: https://scikit-learn.org/stable/modules/model_persistence.html, (visited 18/08/2021).

[30] J. M. B. Cardoso, "Smartfarm: improve sustainability using wireless sensor networks," Master's thesis, Lisboa: Iscte, 2020.

[31] S. Shahidian, R. Serralheiro, J. Serrano, J. Teixeira, N. Haie, and F. Santos, *Hargreaves and Other Reduced-Set Methods for Calculating Evapotranspiration*, 01 2012.

[32] "Edia," [Online] Available: https://www.edia.pt/wp-content/uploads/2019/06/252_preco_agua042017.pdf, (visited 17/09/2021).

[33] "Coeficiente de cultura, kc," [Online] Available: http://www.fao.org/3/x0490e/x0490e0b.htm, (visited 5/05/2021).

Apêndices

APÊNDICE A

Resultados dos Algoritmos na Previsão dos Vários Parâmetros

Na Tabela A1 é possível observar-se os resultados da regressão para o parâmetro da temperatura, usando o dataset diário.

TABELA A1. Resultados da Regressão da Temperatura utilizando Dataset Diário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.476	0.473	0.639
Redes Neuronais	2.936	2.862	2.910
Árvore de Decisão	0.525	0.552	0.727
Regressão Linear	6.172	6.172	6.164

Os modelos que se destacaram pelos melhores resultados foram FA e AD, obtendo um EMA de 0.639 ºC e de 0.727 ºC, respetivamente, estes valores dizem respeito à validação que é a coluna que mostra melhor a precisão. Os algoritmos RN e RL obtiveram piores resultados, 2.910 ºC e 6.164 ºC, respetivamente.

TABELA A2. Resultados da Regressão da Velocidade do Vento utilizando Dataset Diário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.359	0.359	0.472
Redes Neuronais	1.175	1.164	1.166
Árvore de Decisão	0.360	0.401	0.527
Regressão Linear	1.143	1.143	1.146

Na Tabela A2 pode-se observar os resultados referentes à regressão da velocidade do vento, tal como aconteceu com os resultados da temperatura, os melhores algoritmos foram o algoritmo FA obtendo um valor de EMA na ordem dos 0.472 km/h e o algoritmo AD que obteve 0.527 km/h. Os restantes algoritmos mais uma vez demonstraram ser menos precisos, o algoritmo RN obteve 1.166 km/h e o RL obteve 1.146 km/h.

Tabela A3. Resultados da Regressão da Precipitação utilizando Dataset Diário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.239	0.239	0.330
Redes Neuronais	3.480	1.918	1.953
Árvore de Decisão	0.255	0.258	0.324
Regressão Linear	1.969	1.969	1.965

Os resultados relativos à regressão da precipitação, Tabela A3, vão ao encontro dos resultados anteriores, sendo que o algoritmo AD e FA obtiveram os melhores valores de EMA, 0.324 mm e 0.330 mm, respetivamente. Os algoritmos RN e RL obtiveram os piores resultados, 1.953 mm e 1.965 mm, respetivamente.

Tabela A4. Resultados da Regressão da Evapotranspiração utilizando Dataset Diário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.156	0.156	0.194
Redes Neuronais	0.835	0.921	0.811
Árvore de Decisão	0.175	0.187	0.230
Regressão Linear	1.945	1.945	1.945

Através da Tabela A4, foi possível observar que o algoritmo FA obteve o valor de 0.194 mm/dia, o algoritmo RN obteu 0.811 mm/dia, enquanto que o algoritmo AD obteve 0.230 mm/dia e por fim, o algoritmo RL obteve o pior valor, de 1.945 mm/dia.

Tabela A5. Resultados da Temperatura utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.790	0.787	1.043
Árvore de Decisão	1.067	1.067	1.487

Em relação aos resultados para a regressão da temperatura com o novo dataset horário, Tabela A5, pode-se observar que o algoritmo FA obteve um valor de EMA de 1.043 ºC e o algoritmo AD obteve 1.487 ºC.

Tabela A6. Resultados da Velocidade do Vento utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	2.341	2.341	2.988
Árvore de Decisão	1.861	1.863	2.554

Nos resultados da velocidade do vento, Tabela A6, o algoritmo que obteve o menor valor de EMA foi o AD, 2.554 km/h e o algoritmo FA obteve 2.988 km/h.

Tabela A7. Resultados da Precipitação utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.072	0.072	0.096
Árvore de Decisão	0.074	0.082	0.106

Os valores relativos à precipitação, Tabela A7, foram similares, com ligeira vantagem para o algoritmo FA, uma vez que obteve menos 0.010 mm de precipitação no que toca ao erro médio absoluto.

Tabela A8. Resultados da Evapotranspiração utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	0.021	0.021	0.022
Árvore de Decisão	0.026	0.026	0.027

Na Tabela A8, é possível observar-se que os dois algoritmos obtiveram resultados muito parecidos no que toca à evapotranspiração, uma diferença de 0.005 mm/dia.

Tabela A9. Resultados da Humidade Relativa utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	4.798	4.829	5.624

Na Tabela A9, o algoritmo FA foi capaz de obter o valor de EMA de 5.624%

Tabela A10. Resultados da Humidade no Solo utilizando Dataset Horário

Modelo	**Padrão**	**Otimizado**	**Validação**
Floresta Aleatória	8.361	8.389	9.513

Por fim, na Tabela A10, o algoritmo obteve o valor de 9.513% de humidade no solo.

APÊNDICE B

Coeficiente de Cultura e Radiação Incidente

Tendo em conta que existem várias plantações de diferentes tipos de alimentos, teve-se em conta alguns dos mais comuns, como se pode observar na Tabela A1. Nesta tabela é possível ver os valores dos vários coeficientes de cultura referentes a cada um dos alimentos, sendo apresentado o valor do coeficiente inicial, médio e final [**33**]. Uma vez que o valor do coeficiente de cultura varia consoante a fase do crescimento.

Tabela A1. Coeficientes de Cultura

Coeficiente	**Kc (inicial)**	**Kc (média)**	**Kc (final)**
Bróculos		1.05	0.95
Cenouras		1.05	0.95
Espinafres		1.00	0.95
Pepinos		1.00	0.75
Batatas		1.15	0.75
Arroz	1.05	1.20	0.90-0.60

Na Tabela A2 pode-se observar os valores de radiação solar extra-terrestre incidente, estes valores dizem respeito à parte do hemisfério norte e como se pode observar ao longo da tabela, estes dependem da latitude e do mês do ano.

TABELA A2. Radiação Solar Extra-terrestre Incidente

Lat º	**Jan**	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
70	0.0	2.6	10.4	23.0	35.2	42.5	39.4	28.0	14.9	4.9	0.1	0.0
68	0.1	3.7	11.7	23.9	35.3	42.0	38.9	28.6	16.1	6.0	0.7	0.0
66	0.6	4.8	12.9	24.8	35.6	41.4	38.8	29.3	17.3	7.2	1.0	0.1
64	1.4	5.9	14.1	25.8	35.9	41.2	38.8	20.0	18.4	8.5	2.4	0.6
62	2.3	7.1	15.4	26.6	36.3	41.2	39.0	30.6	19.5	9.7	3.4	1.3
60	3.3	0.3	6.5	27.5	36.6	41.2	39.2	31.3	20.6	10.9	4.4	2.2
58	4.3	9.6	17.7	28.4	37.0	41.3	39.4	32.0	21.7	12.1	5.5	3.1
56	5.4	10.8	18.9	29.2	37.4	41.4	39.6	32.6	22.7	13.3	6.7	4.2
54	6.5	12.0	20.0	30.0	37.8	41.5	39.8	33.2	23.7	14.5	7.8	5.2
52	7.7	13.2	21.1	30.8	38.2	41.6	40.1	33.8	24.7	15.7	9.0	6.4
50	8.9	14.4	22.2	31.5	38.5	41.7	40.2	34.4	25.7	16.9	10.2	7.5
48	10.1	15.7	23.3	32.2	38.8	41.8	40.4	34.9	25.6	18.1	11.4	8.7
46	11.3	16.9	24.3	32.9	39.1	41.9	40.6	35.4	27.5	19.2	12.6	9.9
44	12.5	18.0	25.3	33.5	39.3	41.9	40.7	35.9	29.4	20.3	13.9	11.1
42	13.8	19.2	26.3	34.1	39.5	41.9	40.8	36.3	29.2	21.4	15.1	12.4
40	15.0	20.4	27.2	34.7	39.7	41.9	40.8	36.7	30.0	22.0	16.3	13.5
38	16.2	21.5	28.1	35.2	29.9	41.8	40.8	37.0	30.7	23.6	17.5	14.8
36	17.5	22.6	29.0	35.7	40.0	41.7	40.8	37.4	31.5	24.6	18.7	16.1
34	18.7	23.7	29.9	36.1	40.0	41.6	40.8	37.6	31.2	25.6	19.9	17.3
32	19.9	24.8	30.7	36.5	40.0	41.4	40.7	37.9	32.8	26.5	21.1	18.5
30	21.1	25.8	31.4	36.8	40.0	41.2	40.6	38.0	33.4	27.6	22.2	19.8
28	22.3	25.8	32.2	37.1	40.0	40.9	40.4	38.2	33.9	28.5	23.3	21.0
26	23.4	27.6	32.8	37.4	39.9	40.6	40.2	36.2	34.5	29.3	24.5	22.2
24	24.6	28.8	33.5	37.6	39.7	40.3	39.9	38.3	34.9	30.2	25.5	23.3
22	25.7	29.7	34.1	37.8	39.5	40.0	39.6	38.4	35.4	31.0	26.6	24.5
20	26.8	30.6	34.7	37.9	9.3	39.5	39.3	38.3	35.8	31.8	27.7	25.6
18	27.9	31.5	35.2	38.0	39.0	39.1	38.9	38.2	35.1	32.5	28.7	26.8
16	28.9	32.3	35.7	38.1	38.7	38.6	38.5	38.1	36.4	33.2	29.6	27.9
14	29.9	33.1	36.1	38.1	38.4	38.1	38.1	38.0	36.7	33.9	30.6	28.9
12	30.9	33.8	36.5	38.0	38.0	37.6	37.5	37.8	36.9	4.5	31.5	30.0
10	31.9	34.5	36.9	37.9	37.6	37.0	37.1	37.5	37.1	35.1	32.4	31.0
8	32.8	35.2	37.2	37.8	37.1	36.3	36.5	37.2	37.2	35.6	33.3	32.0
6	33.7	35.6	37.4	37.6	36.6	35.7	5.9	36.9	37.3	36.1	34.1	32.9
4	34.6	36.4	7.6	37.4	36.0	35.0	35.3	36.5	37.3	36.6	34.9	33.9
2	35.4	37.0	37.8	37.1	35.4	34.2	34.6	36.1	37.3	37.0	35.6	34.8
0	36.2	37.5	37.9	36.8	34.8	33.4	33.9	35.7	37.2	37.4	36.3	35.6

Devido às unidades do coeficiente estarem expressas em $\mathrm{MJm^{-2}dia^{-1}}$ é necessário proceder à sua conversão para que fique nas unidades mm/dia, tal é possível recorrendo à Equação B.1.

$$R_0[mm/dia] = R_0[MJm^{-2}dia^{-1}] \times \frac{238.85}{597.3 - 0.57 \times temperatura_{média}} \tag{B.1}$$

Na Tabela A3 pode-se observar os parâmetros por defeito usados ao longo da execução dos vários algoritmos no trabalho.

Tabela A3. Parâmetros Configuração por Defeito

Algoritmo	**Configuração**
Floresta Aleatória	n_estimators = 10, max_features = 'auto', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, bootstrap = True, criterion = 'mse', n_jobs = -1, random_state = 'RSEED'
Redes Neuronais	activation = 'redu', solver = 'adam', alpha = '0.0001'
Árvore de Decisão	n_estimators = 10, splitter = 'best', max_features = 'auto', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, bootstrap = True, criterion = 'mse'
Regressão Linear	fit_intercept = True, normalize = True, copy_X = True, n_jobs = -1, verbose = 2, n_iter = 100

APÊNDICE C

Contribuições Científicas

Prediction of Weather Forecast for Smart Agriculture supported by Machine Learning

Francisco Raimundo
Instituto Universitário de Lisboa (ISCTE-IUL)
Lisbon, Portugal
fjsnr@iscte-iul.pt

André Glória
Instituto Universitário de Lisboa (ISCTE-IUL)
Lisbon, Portugal
afxga@iscte-iul.pt

Pedro Sebastião
Instituto Universitário de Lisboa (ISCTE-IUL)
Lisbon, Portugal
pedro.sebastiao@iscte-iul.pt

***Abstract*—This paper introduces a study done to evaluate the use of machine learning regression techniques to predict the weather conditions of agricultural fields for smart irrigation systems. The proposed methodology is able to predict the temperature, precipitation, wind speed and evapotranspiration based on the field location and day. To discover the best model to achieve this, a set of machine learning techniques were implemented, including Linear Regression, Decision Tree, Random Forest and Neural Networks, being the results compared. Results shown that Random Forests and Decisions Trees achieve the best efficiency, after cross-validation. This paper includes a detailed description of the methodology, its implementation and the experimental results.**

***Index Terms*—Machine Learning, Weather Forecast, Smart Agriculture, Internet of Things, Regressions**

I. Introduction

Due to climate change and population growth, nowadays is, more than ever, needed to create ways to produce food sources not only able to feed the population, but that can also help minimize or mitigate the environmental impacts of intensive agriculture [1].

With the rise of technology and the ability of having smart systems controlling everything in our life's, also agriculture started to steer towards this intelligent and autonomous system, with the Internet of Things (IoT) bringing a new way of controlling the entire process, from seeding to harvest.

Although these systems being able to provide a better understanding of the conditions of the fields, the amount of information gathered by them is unbearable of the farmer to analyze and to decide upon them in real time. This creates a new problem, as this intelligence needs to do in real-time and autonomously and this is how Machine Learning (ML) arises.

Machine Learning is a technology that allows machines to learn how to study data in a more efficient and autonomous way [2], allowing for the quicker discovery of patterns, classify or predict data, without human intervention [3].

Machine Learning and IoT are two fields that are more and more dependent on each other, and smart agriculture is one the main areas that benefit from this. With several studies conducted using machine learning techniques it can be applied in the whole spectrum of agriculture, from seeding, harvest, pest and deceases control, quality checks and all the other steps involved in the supply chain of agriculture, as stated by [4] on his review on the application of machine learning in agriculture.

This work was supported in part by ISCTE - Instituto Universitário de Lisboa from Portugal under the project ISCTE-IUL-ISTA-BM-2018

Since agriculture heavily depends on weather forecast to adjust irrigation, harvest and almost every step to the supply chain [5], it becomes critical to have the capability of predicting the weather situation for those particular fields, instead of relying only of weather forecast provided by meteorological companies for that region. Mainly in irrigation, knowing if it will rain in the next hours, can help decide to postpone irrigation. Doing this in an autonomous way can lead to prevent the waste of water and to create more sustainable process in agriculture.

Thus, this paper presents a study of Machine Learning Regression algorithms, where Random Forest (RF), Neural Networks (NN), Decision Trees (DT) and Linear Regression (LR) will be studied, in order to build a system capable of predicting the weather forecast for the upcoming hours, in terms of temperature, wind velocity and direction, precipitation and evapotranspiration, based on the location of the field and historical data. This predictive model, using the best model studied, can in the future be implemented in an agricultural autonomous irrigation system.

After this introduction, some research found in the literature is presented, being followed by an introduction to the Machine Learning techniques. Then, the methodology for the study is presented, including the used dataset and training process. Results and discussion based on the followed methodology are given, with the conclusions from our research being presented at the end.

II. Related Work

With the great growth of Machine Learning associated to smart agriculture, several researches can be found in the literature. A great survey of this research spectrum can be found in [4], covering several applications from seeding, harvest, pest and deceases control, irrigation, quality checks and several others.

In [6], the authors propose a weather forecast model using Linear Regression and a typical equation model. They demonstrate that Machine Learning can be as reliable as

the normal predictive methods for future weather conditions, achieving, for example, a 0.599866 degree margin of error when predicting temperature. Using the same methodology but with Neural Networks and Multiple Linear Regression, in [7], the authors were able to predict future temperature conditions, achieving a margin of error of 1.0782 degrees. Several works were also found using Linear Regressions to predict the weather for the next day, achieving precisions of 90% [8] and 84% [9], when predicting temperature and precipitation.

Not only regressions were used for weather forecasting, in [10], the authors developed a classification computational model based on Decision Trees to predict the next day weather, achieving an accuracy of 82.62%.

The methodology presented in this paper will differ from the work found in the literature in the way that will explore more regression models, other than Linear Regressions. Also, a larger dataset will be used, instead of only a few days, as in the some of the presented works. Not only this will allow us to study how other models work in predicting weather forecast, it will also be helpful to compare to the previous results and understand the importance of having deeper previous knowledge.

III. Machine Learning Techniques

Machine Learning has various learning techniques that can be used, with the most common being supervised learning and unsupervised learning. For supervised learning, a pair of input-output training data is provided with this technique analyzing the data and the relationship between them and in the end, being possible to predict a function from the input with the best estimation of output [11]. A supervised learning algorithm learns from labeled training data, helping to predict outcomes for unforeseen data.

Supervised learning can be divided into regression and classification [11]. Regressions are a predictive analysis that correlates the final value with a set of variables that do not depend on any value, being independent [12]. For this particular case study, regressions are important since the values are not defined in a small range of values or in a binary way, being impossible to predict using classification [13].

Inside supervised learning there are multiple techniques that can be used to do regressions, being the most commons described next.

1) Random Forest (RF) is created by a large set of random decision trees, where each individual tree is responsible for dividing the data in a recursive way, in order to predict the best result. The final result is based on the average results for all the trees [14].
2) Neural Network (NN) work in the fashion as the human brain, being composed by a set of neurons that are interconnected. These neurons analyze the information of the entry data and send their output to the next neurons, until a valid result is achieved [15]. The Multilayer Perceptron (MLP) are a variation of NN, composed by multiple neurons between the entry and output layer, being adjustable as needed for the complexity of the needed output [16].
3) Decision Trees (DT) is a tree-based algorithm based on nodes and leafs, in which the goal is to reach a pure leaf. The data division is done based on the parameters found on the node, creating multiple sub-nodes until a leaf is reached, where no more division is needed [17].
4) Linear Regression (LR) allows for the model to create a hyperplane based on the input data, being this line able to predict the data based on the inputs. This line is calculated with previous data and is able to predict the outcome of new data [17].

IV. Regression Model

As said before, our methodology uses a regression model for the computation analysis. For this model to work, it needs to be trained with previously knows data and configured to achieve the best accuracy and lowest error possible. The next sections explain, in detailed, how this process was done and the obtained model results.

The goal of this model is to receive as input the location of the field (Latitude and Longitude) and the day, month and year and output the predicted weather forecast for temperature, wind speed, precipitation and evapotranspiration.

A. Dataset

For a machine learning model to work, it needs to be trained with a set of supervised data, that include the output desired for a set of parameters. It is with these data, that the model will learn and predict upon future data.

The used dataset is composed of the "JRC MARS Meteorological Database containing meteorological observations from weather stations interpolated on a 25x25 km grid, on a daily basis from 1979 to the last calendar year completed, for the European Union and neighbouring countries", provided by the Agri4Cast database [18].

This allows to have a dataset of multiple weather parameters from the last 40 years divided by cities, districts or countries, across the European Union. In this case, the data from Central Alentejo, in Portugal, was used, since it is a region mainly used for agriculture. With that in mind, the final dataset is composed of 383526 entries, with the following parameters:

- Latitude – Latitude of the field;
- Longitude – Longitude of the field;
- Altitude – Altitude of the field;
- Day – Day of the observation;
- Month – Month of the observation;
- Year – Year of the observation;
- Temperature_Max – Maximum temperature registered [ºC];
- Temperature_Min – Minimum temperature registered [ºC];
- Wind Speed – Average wind speed registered [km/h];
- Precipitation – Total precipitation registered [mm];
- Evapotranspiration – Evapotranspiration registered [mm/day];

B. Training Methodology

In order to train the regression model, the following steps were done, using Python, the scikit-learn libraries [19] and the Anaconda environment.

1) For each algorithm a model was trained using the corresponding dataset and the default configuration parameters. This allowed for quick comparison of the performance, in terms of both accuracy and margin of error, of each model and understand which are more likely to guarantee best results and which need to be improved to achieve them. The scikit-learn, an open source Machine Learning library developed for Python implementation [19], was the selected framework for the development of the machine learning models.
2) The obtained model for each algorithm is submitted into a hyper parametrization tuning, that compares the model performance using different model configurations parameters, to understand which is the configuration that obtains the best performance, facing the dataset and the goal. For this, a method provided by scikit-learn called RandomizedSearchCV was used, which performs the fit and training of the algorithm under study, calculating which parameters are best suited to it [20];
3) To guarantee that the model is stable, after finding the best configuration and train the model, a Stratified K-Fold cross validation is performed, in order to guarantee that the model is not under or over fitted. Using five folds, it is possible to use a different set of training and validation data on each fold, allowing for the model to check on every single datapoint. This way, it is possible to really understand the model performance, as each of the folds will produce a result, that is averaged at the end, allowing for a reduced error margin and variation, as more data is used to fit the model;

V. Results & Discussion

The presented training methodology was followed in order to obtain the best model possible to predict the weather forecast, based on field location and day.

To train, validate and test the model, the presented dataset will be used, being divided into three groups: 70% for training, 20% for validation and 10% for testing. To evaluate the model performance, and since a regression is used, the Mean Absolute Error (MAE) metric will be used, as it is the most common metric for regression. It measures the average absolute error between the real data and the estimated value, using (1) [21], where P_{rx} is the real value, $\hat{P}_{rx}$ is the estimated value, and N the number of samples.

$$MAE[dBm] = \frac{1}{N} . \sum_{i=1}^{N} |P_{rx_i} - \hat{P}_{rx_i}| \quad (1)$$

The estimated data nearly matched the real data when MAE is near 0.

The results of the regression models to predict the temperature, averaged for the maximum and minimum scenario, for each step of the presented methodology, are displayed on Table I and allows us to conclude which is the best model to use in this scenario.

TABLE I
Temperature Regression Results

Model	Default	Optimized	Validation
Random Forest	0.476	0.473	0.639
Neural Networks	2.936	2.862	2.910
Decision Tree	0.525	0.552	0.727
Linear Regression	6.172	6.172	6.164

As is possible to see, the values have a noticeable variation depending on the stage of the methodology. It can be seen that the hyper parametrization, following the expected behavior due to the presented methodology, since after the first test with the default values, these values will be tuned to improve, being predictable that in the optimized scenario better results will be obtained. Then, in the cross-validation stage, as multiple combinations of the dataset are tested and the MAE for each fold is averaged, it is expected that the MAE increase. With this in mind, only the validation results will be analyzed, since they are the ones that best show the behavior of the model.

Comparing the used regression models, it is possible to check that Random Forest and Decision Trees have similar results, being Random Forest the best one, achieving 0.639 degrees of error, minus 0.158 degrees than Decision Trees. As for Neural Networks and Linear Regression, the Linear Regression achieved the worst results, with a MAE almost 6 degrees higher than Random Forest.

Comparing to the works presented on the literature review, it is possible to check that in [6] they used Linear Regression with a MAE of 0.599, whereas our results were 6.164 using Linear Regression, and in [7], they achieved a MAE of 1.0782 using Neural Networks, with our research achieving 2.910. This shows that the use of different datasets, mainly in terms of number of entries may affect the final results when predicting the same parameter.

Applying the same methodology for the wind speed, the obtained regression results are displayed on Table II and allows us to conclude which is the best model to use in this scenario.

TABLE II
Wind Speed Regression Results

Model	Default	Optimized	Validation
Random Forest	0.359	0.359	0.472
Neural Networks	1.175	1.164	1.166
Decision Tree	0.360	0.401	0.527
Linear Regression	1.143	1.143	1.146

As in the temperature scenario, here the optimized value are also the ones with the lowest MAE, and the validation the ones that best show the model accuracy. Comparing the used models, as in the temperature scenario, also Random Forest and Decision Trees have similar results, being Random Forest the best model, with a MAE of 0.472 km/h, being only 0.055 km/h better than Decision Trees. In this scenario, Neural

Networks and Linear Regression achieved similar results, being Neural Networks the worst scenario.

As for precipitation, the results for each step of the presented methodology, are displayed on Table III and allows us to conclude which is the best model to use in this scenario.

TABLE III
PRECIPITATION REGRESSION RESULTS

Model	Default	Optimized	Validation
Random Forest	0.239	0.239	0.330
Neural Networks	3.480	1.918	1.953
Decision Tree	0.255	0.258	0.324
Linear Regression	1.969	1.969	1.965

Once again, Decision Trees and Random Forest have similar results, but contrarily to the other two scenarios, to predict precipitation, Decision Trees presented the best results, with a MAE of 0.324 mm, only 0.006 mm higher than Random Forest. As in the wind speed, Neural Networks and Linear Regression present similar results, with Linear Regression being slightly worst.

Comparing to the results obtained by the authors of [8], that used Linear Regression with a MAE of 0.536 mm, our results shows that using other models and a more complete dataset, a better result can be achieved.

Finally, the results of the regression models to predict the evapotranspiration, for each step of the presented methodology, are displayed on Table IV and allows us to conclude which is the best model to use in this scenario.

TABLE IV
EVAPOTRANSPIRATION REGRESSION RESULTS

Model	Default	Optimized	Validation
Random Forest	0.156	0.156	0.194
Neural Networks	0.835	0.921	0.811
Decision Tree	0.175	0.187	0.230
Linear Regression	1.945	1.945	1.945

As for the previous models, Random Forest and Decision Tree have similar results, being once again Random Forest the best model, with a MAE of 0.194 mm, 0.036 mm better than Decision Trees. In this scenarios, Linear Regression was, as in the temperature scenario, the worst model by a big margin, almost 2 mm.

With these results it is possible to conclude that the best models to predict future weather conditions based on historical data and field location are Random Forest and Decision Trees, achieving similar results in all scenarios, with Random Forest being the best in three of them. As for Neural Networks and Linear Regressions, the obtained results were never close to the best models, so they will not be used in future applications.

VI. CONCLUSIONS

The goal of this paper was to develop a study of some of the most common algorithms in Machine Learning to predict the weather conditions of an agricultural field, based on its location and day.

It started by studying some related work, finding some research of Machine Learning in weather forecasting with some promising results using Linear Regressions.

A detailed study was made of the Machine Learning techniques and the algorithms used, these being Random Forest, Neural Networks, Decision Tree and Linear Regression. A methodology was followed to do this study, starting with using the default configuration values, then performing hyper parameterization of these values, and finally a cross-validation.

After a careful study of the mean absolute error (MAE) values of all scenarios, it was observed that the Random Forest and Decision Tree algorithms had the best results, achieving the lowest MAE in the optimized and cross validation stages for all the weather conditions. Random Forest achieved best results in Temperature, Wind Speed and Evapotranspiration, with a MAE of 0.639 degrees, 0.472 km/h and 0.194 mm, respectively, and Decision Tree was the best to predict precipitation, with a MAE of 0.324 mm. It was also possible to conclude that Neural Networks and Linear Regressions had the worst results, something that was not expected, since they were the most used models in the literature.

With the best model selected, the next steps in our research are to develop an algorithm that implements this model and automatically decides the best irrigation periods based on weather forecast and then implement that solution in a device and test it in a real case scenario.

REFERENCES

[1] J. Binas, L. Luginbuehl, and Y. Bengio, "Reinforcement Learning for Sustainable Agriculture," in *ICML 2019 Workshop Climate Change: How Can AI Help?*, 2019.

[2] B. Mahesh, "Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Self Flowing Generator View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review," *International Journal of Science and Research*, pp. 381–386, 2018.

[3] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors (Switzerland)*, vol. 18, 8 2018.

[4] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Computers and Operations Research*, vol. 119, 7 2020.

[5] P. P. Pawade and A. S. Alvi, "A Survey on Applications of Machine Learning in Agriculture," *International Research Journal of Engineering and Technology*, 2020.

[6] S. Gupta, K. Indumathy, and G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques," *International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1490–1493, 2016.

[7] T. Anjali, K. Chandini, K. Anoop, and V. L. Lajish, "Temperature Prediction using Machine Learning Approaches," in *2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies : ICICICT-2019*, 2019.

[8] T. R. Anandharajan, G. A. Hariharan, K. K. Vignajeth, R. Jijendiran, and Kushmita, "Weather Monitoring Using Artificial Intelligence," in *Proceedings - International Conference on Computational Intelligence and Networks*, vol. 2016-January, pp. 106–111, Institute of Electrical and Electronics Engineers Inc., 8 2016.

[9] G. Verma, P. Mittal, and S. Farheen, "Real Time Weather Prediction System Using IOT and Machine Learning," in *6th International Conference on Signal Processing and Communication (ICSC)*, vol. 116, pp. 1071–1085, Springer, 1 2020.

[10] Z. Khan, A. W. Khan, U. Mardan, M. Hayat, and Z. U. Khan, "Hourly Based Climate Prediction Using Data Mining Techniques by Comprising Entity Demean Algorithm," *Middle East Journal of Scientific Research*, vol. 21, no. 8, pp. 1295–1300, 2014.

[11] D. Praveen Kumar, T. Amgoth, and C. S. R. Annavarapu, "Machine learning algorithms for wireless sensor networks: A survey," *Information Fusion*, vol. 49, pp. 1–25, 9 2019.

[12] R. N. Behera, K. Das, B. Tech, and A. Professor, "A Survey on Machine Learning: Concept, Algorithms and Applications," *International Journal of Innovative Research in Computer and Communication Engineering*, 2017.

[13] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021.

[14] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for Sustainable Development*, vol. 36, 6 2016.

[15] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 945–949, 2018.

[16] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, p. 26, 2016.

[17] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019.

[18] Agri4Cast, "Gridded Agro-Meteorological Data in Europe," 2021. [Dataset] European Commission, Joint Research Centre (JRC). https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx?o=d.

[19] scikit-learn, "scikit-learn," 2021. [Online] Available: https://scikit-learn.org/stable/, (visited 16/02/2021).

[20] scikit-learn, "RandomizedSearchCV," 2021. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html, (visited 16/02/2021).

[21] D. F. Fernandes, A. Raimundo, F. Cercas, P. J. Sebastiao, R. Dinis, and L. S. Ferreira, "Comparison of Artificial Intelligence and Semi-Empirical Methodologies for Estimation of Coverage in Mobile Networks," *IEEE Access*, vol. 8, pp. 139803–139812, 2020.