APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING PARA CLASSIFICAÇÃO DA APTIDÃO DOS SOLOS PARA O REGADIO

APPLICATION OF MACHINE LEARNING TECHNICS FOR EVALUATION OF THE SOILS CAPABILITY TO IRRIGATION

Pedro Torres, Instituto Politécnico de Castelo Branco, Escola Superior de Tecnologia, Castelo Branco – Portugal; SYSTEC - Research Center for Systems & Technologies, 4200-465 Porto, Portugal. ORCID: https://orcid.org/0000-0003-4835-5022

António Canatário Duarte, Instituto Politécnico de Castelo Branco, Escola Superior Agrária, Castelo Branco – Portugal; Centro de Estudos CERNAS-IPCB, Castelo Branco – Portugal. ORCID: https://orcid.org/0000-0002-0319-378X

João Geraldes, Instituto Politécnico de Castelo Branco, Escola Superior de Tecnologia, Castelo Branco – Portugal;

Sílvia Marques, Instituto Politécnico de Castelo Branco, Escola Superior Agrária, Castelo Branco – Portugal;

Data de submissão: 19/07/2022

RESUMO: Este trabalho consiste no desenvolvimento e validação de modelos de *Machine Learning* para a otimização de um sistema de rega de precisão utilizando algoritmos de classificação. A finalidade é atribuir a cada solo, localizado a sul do concelho do Fundão, Portugal, uma classe de aptidão para o regadio, classes essas que identificam as zonas regáveis, não regáveis bem como as que precisam de intervenção para serem regadas. Os dados dos casos de estudo foram anteriormente recolhidos por uma aluna de Mestrado da Escola Superior Agrária do IPCB (Portugal), onde incluíam vários condicionalismos (características dos solos que podem condicionar a aptidão para o regadio). A análise exploratória dos dados permitiu utilizar apenas os valores dos resultados relativamente às características dos solos que podem condicionar a aptidão para o regadio rejeitando assim todo o cálculo efetuado para a obtenção dos mesmos. Desta forma os dados do caso de estudo foram enriquecidos com esta informação para a aplicação nos algoritmos de *Machine Learning*. Em geral, o facto de retirar estas características que não revelavam impacto no estudo ajudaram a melhorar os modelos de classificação bem como a sua precisão.
Diferentes algoritmos de *Machine Learning* foram desenvolvidos, testados e validados, tais como, *Support Vetor Machine*, *kNN*, Árvore de Decisão, *Naive Bayes* e Regressão Logística, para otimizar um sistema de rega de precisão de modo a atribuir uma a classe de aptidão de rega a novos solos introduzidos. A

comparação dos modelos demonstrou que o método *Naive Bayes* é o que apresenta uma melhor precisão na altura de gerar uma classe de previsão.

PALAVRAS-CHAVE: Aptidão solos regadio, Machine Learning, Scikit-Learn, Aprendizagem Supervisionada.

ABSTRACT: This work consists of the development and validation of Machine Learning models for the optimization of a precision irrigation system using classification algorithms. The purpose is to assign to each soil, located in the south of the municipality of Fundão, Portugal, an class of capability to irrigation, classes that identify the irrigable and non-irrigated areas as well as those that need intervention to be irrigated. Data from the case studies were previously collected by a Master's student at the Escola Superior Agrária – IPCB (Portugal), which included several constraints (characteristics of soils that may affect the suitability for irrigation). The exploratory analysis of the data allowed us to use only the values of the results regarding the characteristics of the soils that may affect the suitability for irrigation, thus rejecting all the calculation made to obtain them. In this way, the case study data were enriched with this information for application in Machine Learning algorithms. In general, removing these features that had no impact on the study helped to improve the classification models as well as their accuracy.
Different Machine Learning algorithms were developed, tested, and validated, such as Support Vector Machine, kNN, Decision Tree, Naive Bayes and Logistic Regression, to optimize a precision irrigation system in order to assign an irrigation suitability class. to new introduced soils. The comparison of the models showed that the Naive Bayes method is the one that presents the best precision when generating a prediction class.

.

KEY-WORDS: Soils capability irrigation, Machine Learning, Scikit-Learn, Supervised Learning.

1. INTRODUÇÃO

Nos dias de hoje o conceito de Agricultura Inteligente ou de Precisão tem se destacado e é fortemente enriquecido com a aplicação de sensores que possibilitam a avaliação e monitorização das condições ambientais, bem como a análise inteligente dos dados produzidos por estes sensores. Este conceito está associado à utilização de equipamentos de alta tecnologia com algoritmos de precisão que tornaram mais eficiente e eficaz este conceito. A agricultura de precisão possibilita que o agricultor “olhe” para uma parcela de solo e perceba que a deve tratar de forma diferenciada, esta nova mentalidade veio revolucionar a forma de ser agricultor. Duas máximas estão associadas a esta mentalidade,

o aumento do rendimento da economia dos agricultores e a redução do impacto ambiental causado pelo setor agrícola. Um dos mecanismos impulsionadores deste conceito é o *Machine Learning (ML)* [1], que consiste em fazer com que as máquinas aprendam através de experiências que lhes são fornecidas sem ser propriamente necessário programá-las. Acoplando a este conceito outras tecnologias como Big Data e computação de alto desempenho, novas oportunidades foram desenvolvidas para quantificar e compreender processos intensivos de dados em ambiente agrícola.

O ML aplicado ao setor agrícola utiliza um conjunto de modelos de aprendizagem bem definidos que recolhem dados específicos bem como aplicam algoritmos para obter resultados esperados. Os modelos de ML podem ser usados para prever a qualidade do solo, a quantidade de água necessária para regar, entre outros.

1.1. Enquadramento

Neste trabalho, aplica-se, avalia-se, prevê-se, classificam-se e validam-se a aptidão dos solos para rega, através de modelos de precisão com o auxílio a técnicas de ML (*Machine* Learning). Os diferentes modelos aplicados visam estabelecer uma comparação entre diferentes métodos de classificação e avaliar a sua precisão. O objetivo passa por avaliar a aptidão para o regadio dos solos existentes a sul do concelho do Fundão, Portugal, com a intenção de expandir o regadio da Cova da Beira a zonas onde não existe regadio.

O trabalho tem uma aplicabilidade prática no qual o torna motivante para o seu desenvolvimento visto que a importância do regadio é essencial para sustentar as atividades agrícolas e pela qual dependem muitas famílias de agricultores. O regadio da Cova da Beira não é extenso o suficiente para que outras zonas menos beneficiadas e com grande potencial agrícola se possam servir dele. O estudo da aptidão dos solos para o regadio pode determinar quais as áreas com capacidade para serem regadas, ou não.

Com o aumento da quantidade de dados recolhidos sobre as características do solo estão reunidas as condições para aplicar técnicas de *Machine Learning* no setor agrícola. O grande desafio passa por encontrar o algoritmo que melhor se aplica ao objetivo do projeto. Os diferentes algoritmos

de ML que consigam modelar e prever uma classe de aptidão de rega podem ser importantes para ajudar a perceber quais as áreas que podem ser regadas, com o objetivo de estender o regadio a zonas menos beneficiadas e com forte potencial agrícola. Os algoritmos de ML como também a sua aplicabilidade estão em constante expansão. Estes algoritmos categorizam-se em diferentes tipos de aprendizagem (supervisionada, não supervisionada e de reforço) bem como também por técnicas de formulação (classificação e regressão). A seleção de um algoritmo depende muito do tipo de problema que se vai abordar. Na rega de precisão a técnica de classificação é bastante utilizada para prever uma dotação correta de quando regar ou não regar, devido ao facto de as características dos solos serem apresentadas em forma de dados, onde fica presente a informação sobre as características dos solos bem como a avaliação de cada solo.

1.2. Inteligência Artificial e *Machine Learning*

Atualmente Inteligência Artificial (I.A.) é um conceito com bastante aplicabilidade, caracteriza-se como uma ciência ou ramo da engenharia, que procura estudar e compreender o fenómeno da inteligência e por outro lado compreender o modo como os seres humanos pensam, a fim de modelar o pensamento em processos computacionais e consequentemente construir um corpo de explicações algorítmicas dos processos mentais humanos.

Machine Learning é um sub ramo da IA, onde se usam algoritmos para adquirir dados, inferir com eles, e fazer uma determinação ou previsão sobre algo. A máquina é “treinada” através de uma grande quantidade de dados e implementa algoritmos que lhe dão a habilidade de aprender como executar a tarefa [2]. O *Machine Learning* ensina os computadores a fazer o que os humanos e os animais naturalmente fazem, aprender através de experiências. Os algoritmos de *Machine Learning* utilizam métodos de aprendizagem computacional para “aprender” a informação através de dados sem depender de equações predeterminadas como modelos. Os algoritmos melhoram de forma adaptativa a sua performance à medida que o número de amostras disponíveis aumenta.

1.3. Aprendizagem Supervisionada e Não Supervisionada

Quando se fala em algoritmos de ML conduz, normalmente, à referência de dois paradigmas: aprendizagem supervisionada e não supervisionada[3]. Na aprendizagem supervisionada são usados dois conjuntos de dados, o de input e output esperado. Na aprendizagem não supervisionada apenas é apresentado um conjunto de dados, input.

A aprendizagem supervisionada está diretamente relacionada com a previsão enquanto a aprendizagem não supervisionada relaciona padrões num conjunto de dados de modo a agrupá-los [3]. O objetivo da aprendizagem supervisionada é construir um modelo que faça previsões baseadas em evidências na presença de incertezas. O algoritmo de aprendizagem supervisionada usa um conjunto de dados de entrada (inputs) e respostas (outputs) conhecidos e treina o modelo para gerar previsões razoáveis para a resposta de novos dados. Todas as técnicas de aprendizagem supervisionada são sob a forma de classificação que preveem respostas discretas ou regressão para previsão de respostas contínuas.

A regressão é habitualmente utilizada para a previsão de valores de variáveis dependentes (variáveis que se pretende prever) a partir de uma ou mais variáveis independentes (atributos conhecidos) e nos casos em que essas mesmas variáveis são contínuas. Trata-se de uma tarefa utilizada na aproximação dos dados recebidos.

- A classificação consiste no processo de encontrar um modelo que descreva e distinga classes de dados ou conceitos. Depois de encontrado esse modelo, é possível aplicá-lo de forma a prever a classe de um novo objeto. O modelo gerado é baseado na análise de um conjunto de dados, designado por conjunto de treino.

Para a execução da tarefa de classificação é possível aplicar uma série de métodos de aprendizagem automática nomeadamente: árvores de decisão, regras de classificação (regras if-then), programação lógica indutiva, SVM, redes bayesianas, entre outros.

- A aprendizagem não supervisionada procura encontrar padrões similares entre as várias características dos dados. *O Clustering* é a técnica mais

comum de aprendizagem não supervisionada, é usada para análise exploratória de dados para encontrar padrões e agrupar os dados em grupos, clusters.

A Aprendizagem por reforço [4] foca-se no objetivo de maximizar a recompensa final, ou seja, atingir o objetivo definido com o máximo número de recompensas acumuladas em cada passo e ação, evitando dessa forma as ações com resultados negativos e obtendo a solução ótima para todo o problema, sem qualquer ajuda. A **Figura *1*** esquematiza o modelo de aprendizagem por reforço.

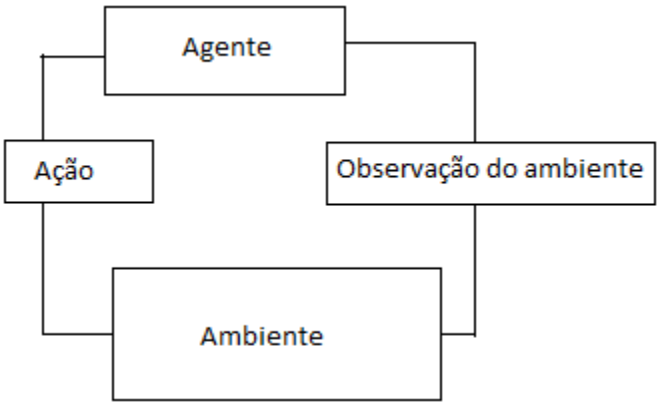

Figura 1 - Modelo de aprendizagem por reforço.

1.4. Algoritmos de Classificação (abordados neste trabalho)

No estado da arte existem diversos algoritmos de classificação e cada algoritmo tem abordagens diferentes de aprendizagem.

O melhor método ou o método único não existe. Encontrar o algoritmo certo, em parte, é ir por tentativa erro – os cientistas altamente experientes não conseguem provar se o método irá funcionar sem o experimentarem. Mas a seleção dos algoritmos também depende do tamanho e tipo de dados com os quais estamos a trabalhar.

1.4.1. Método 1, kNN (K-nearest neighbors)

O kNN [5] categoriza objetos baseando-se na proximidade das observações,assume que objetos próximos são semelhantes, logo pertencem à mesma classe. Usa métricas como a distância Euclidiana, distância de

Hamming, distância de Manhattan e Distância de Markowski para encontrar o vizinho mais próximo.

1.4.2. Método 2, SVM (Support-vector machine)

O Support Vector Machine (SVM) [6] é um método de aprendizagem supervisionada, utilizado quer para classificação como para regressão.

Em tarefas que requerem a aprendizagem de duas classes, o objetivo do SVM é encontrar a melhor função de classificação que permita a distinção entre membros de duas classes num conjunto de treino . As SVM's classificam os dados de modo a encontrar uma função de classificação linear que separa os dados por um hiperplano que atravessa as duas classes. O melhor hiperplano para uma SVM tem de apresentar uma margem larga entre as duas classes, quando os dados estão linearmente separados. Se os dados não estiverem separados linearmente, uma função de perda é utilizada.

1.4.3. Método 3, Regressão Logística

O método de Regressão Linear [7] pode ser utilizado para estudar a relação entre duas variáveis . Por ser um método simples, a regressão logística é usada como um ponto de partida para problemas de classificação binária (MathWork em Introducing Machine Learning). A expressão linear que explica a relação binária é dada por:

$$y = B_0 + B_1 x + u$$

Onde,

y, consiste na variável dependente;
x, diz respeito á variável independente;
u, ou erro, fatores que influenciam y para além de x;
B0, parâmetro de interceção, conhecido por constante;
B1, representa o declive na relação entre x e y.

Para estimar B0 e B1 é necessário recorrer ao método dos mínimos quadrados.

1.4.4. Método 4, Naive Bayes

O classificador Naive Bayes [8] baseia-se na aplicação do teorema Bayes, expressão matemática usada para o cálculo da probabilidade de um dado evento acontecer visto que outro ocorreu, e assume que a presença de características

particulares numa classe não está relacionada com as características de outra classe. O desempenho deste método pode ser comparado ao método de Árvores de Decisão . O método classifica novos dados baseados na probabilidade de pertencerem a uma determinada classe . O Teorema de Bayes é expresso matematicamente pela seguinte equação:

$$P(A,B) = \frac{P(B,A)P(A)}{P(B)}$$

Onde,

P(A) e P(B), são as probabilidades de A ocorrer e B ocorrer;

P(A,B) é a probabilidade de A acontecer dado que B ocorreu;

P(B,A) é a probabilidade de B acontecer dado que A ocorreu.

1.4.5. Método 5, Árvores de Decisão

As Árvores de Decisão [9] têm por base algoritmos que dividem o conjunto inicial de dados em subconjuntos mais homogéneos que por sua vez se podem dividir em subconjuntos ainda mais homogéneos . Uma árvore de decisão é formada por um conjunto de nós de decisão, perguntas, que permitem a classificação de cada caso. A árvore de decisão permite prever as respostas dos dados seguindo uma rota desde o início ao fim do último nó. A árvore consiste em condições de ramificação em que o valor da previsão é comparado a um peso de treino. O número de ramificações e os valores dos pesos são determinados no processo de treino. Modificações adicionais podem ser usadas para simplificar o modelo.

As árvores de decisão caracterizam-se por utilizarem a estratégia de divisão e conquista. Sendo assim, focam-se num problema considerado complexo, dividindo-o em problemas mais simples e recursivamente aplicando a mesma estratégia a sub-problemas. No final, as soluções dos sub-problemas podem ser combinadas para gerar a solução do problema inicial.

3. CASO DE ESTUDO – Classificação da aptidão dos solos para o regadio na região da Cova da Beira

A agricultura é afetada pelos problemas de seca e escassez de água em algumas regiões do País, imediatamente se percebe a importância da estratégia de implementar um regadio para sustentar níveis de produtividade e redução de

custos nos processos agrícolas de maneira a não comprometer as gerações futuras.

Segundo a Direção Geral de Agricultura (DGA), Portugal apresenta níveis de precipitação média anual da ordem dos 700mm, contudo o desequilíbrio da precipitação pelo país gera problemas de escassez principalmente nos meses de abril a setembro. Posto isto o desenvolvimento da cultura vegetativa no período primavera-verão torna-se difícil.

O regadio surge então com um papel preponderante no que diz respeito à sustentabilidade da agricultura como também para o desenvolvimento socioeconómico das zonas rurais. É importante realçar que o uso de regadio não implica que os utilizadores tomem medidas eficientes para assegurar a quantidade de água que cada um necessita, dado que este recurso (água) em questões económicas, sociais e ambientais é deveras importante.

Na região da Cova da Beira, o regadio abrange vários concelhos: Sabugal, Belmonte, Penamacor, Covilhã e Fundão, com uma área de 12360ha, [10]. Atualmente está em estudo o alargamento do regadio, concretamente a sul da Serra da Gardunha, para zonas com forte potencial na atividade frutícola onde se verificou um aumento das áreas de pomar.

Dada a importância do regadio no que diz respeito ao setor agrícola, a analise e avaliação da aptidão que os solos têm para receber água é crucial para a construção do mesmo. A aptidão dos solos para a atividade agrícola permite avaliar a capacidade de cada parcela de solo para uma determinada cultura com o objetivo de tornar rentável os recursos bem como aumentar a produtividade.

3.1. Análise dos Dados

Nesta secção são descritos os condicionantes dos solos, estes apresentam as características dos solos utilizadas para efeitos de estudo que irão permitir identificar uma classe de aptidão de rega. Cada parcela de solo foi classificada relativamente a diversas características tais como, a natureza do solo (NR), como sendo boa, regular, sofríveis, medíocres e maus.

A espessura efetiva do solo (E), para espessura superior a 100 cm, classe E1, classe E2 para espessuras entre 60 cm e 100 cm, E3 para espessuras entre os

40 cm e 60 cm, classe E4 para espessuras entre os 25 cm e 40 cm e classe E5 para espessuras de solo inferiores a 25 cm.

A capacidade de água utilizável (CA), por uma camada de solo, corresponde a mais um critério da classificação para a aptidão de rega e estabelece 5 classes. A capacidade de água utilizável pelas plantas mais elevadas, requerem regas mais alargadas e por consequência menos mão-de-obra.

Condições de drenagem (HD), muitas culturas precisam de boas condições de drenagem dos solos para que consigam produzir com maior eficiência e eficácia. Um solo com fraca capacidade de escoar e eliminar o excesso de água pode alagar-se. Esta característica compreende 5 classes. Solos com boa capacidade de drenagem (HD1) até uma fraca capacidade de drenagem (HD5).

Outra característica é o risco de erosão (RE), as práticas agrícolas e a forma inadequada como se utiliza o solo são os principais fatores responsáveis pelo processo de erosão, como consequência perdem-se camadas superficiais reduzindo assim a sua produtividade.

O risco de inundação (HI), esta característica é importante no sentido em que zonas com elevado risco de inundação não necessitem de ser regadas, este condicionalismo apresenta 5 classes, (HI1) para solos com riscos nulos de inundação e (HI5) até riscos elevados.

A perigosidade (P), é simplesmente a característica, que afeta o uso de máquinas agrícolas. Define-se em 5 classes, (P1) para solos sem pedregosidade até (P5) que impossibilitam o uso de máquinas agrícolas.

Por último a salinidade (S) dos solos, que permite identificar o grau de afetação dos solos nas culturas. Esse impacto é definido em 5 classes, na classe (S1) são incluídos os solos que não afetam qualquer tipo de cultura, contudo a classe (S5) estão incluídos os solos que impedem o desenvolvimento das culturas.

Com base nestes condicionalismos a tabela seguinte faz corresponder as características de cada parcela de solo a uma classe de aptidão de rega.

Tabela 1 – Condicionalismos e classe de aptidão do solo para o regadio.

CONDICIONALISMOS (Características dos solos que podem condicionar a sua aptidão para o regadio)	**CLASSES DE APTIDÃO DO SOLO PARA O REGADIO**						
	CLASSE I	CLASSE II	CLASSE III	CLASSE IV	CLASSE V	CLASSE VI	CLASSE VII
NATUREZA DO SOLO (**NR**)	NR1	NR2	NR3	NR4	NR5	NR5	NR5
ESPESSURA EFECTIVA (**E**)	E1	E2	E3	E4	E5	E5	E5
RISCOS DE EROSÃO (**RE**)	RE1	RE2	RE3	RE4	RE5	RE5	RE5
CAPACIDADE DE ÁGUA UTILIZÁVEL (**CA**)	CA1	CA2	CA3	CA4	CA5	CA5	CA5
DRENAGEM (**HD**)	HD1	HD2	HD3	HD4	HD5	HD5	HD5
RISCOS DE INUNDAÇÃO (**HI**)	HI1	HI2	HI3	HI4	HI5	HI5	HI5
PEDREGOSIDADE E AFLORAMENTOS ROCHOSOS (**P – R**)	P1 – R1	P1 – R1	P2 – R2	P3 – R3	P4 – R4	P4 – R4	P5 – R5
SALINIDADE E/OU ALCALINIDADE (**S**)	S1	S2	S3	S4	S5	S5	S5

3.2. Resultados de Classificação

Os dados são estruturados da seguinte maneira: o ficheiro Excel contém 696 linhas de dados. Para efeitos de programação dos métodos, 75% dos mesmos são utilizados para treinar o modelo enquanto os restantes 25% são utilizados para testar o modelo de previsão. Para gerar uma classe de previsão baseada no histórico os algoritmos obrigatoriamente precisam de uma nova entrada, para tal utilizaram-se os mesmos dados em todos os modelos para avaliar a precisão dos modelos, [3,5,4,5,2,3,1]. Recorreu-se à Matriz de Confusão para avaliar os modelos de classificação, ou seja, se o mesmo previu de forma correta a classe desejada. De uma forma muito simples a Matriz de Confusão é uma tabela que

mostra as frequências de classificação para cada classe do modelo. Os valores registados da precisão estão de acordo com a nova entrada.

A fórmula seguinte permitiu validar a precisão dos modelos.

$$Precisão = \frac{\text{TP}}{(TP+FP)} \quad (1)$$

TP: True Positives, FP: False Positives

Tabela 2 – Comparação de métodos.

Método	Métrica de precisão
kNN	93.86 %
SVM	92.91 %
RL	90.80 %
NB	94.63 %
AD	94.44 %

3.2.1. Método kNN

O Método kNN, apresentou uma precisão de 93.8% e atribuiu ao novo dado introduzido a classe de previsão, 5. Segundo a matriz de confusão da Figura 9, as 19 amostras de classe 3, foram corretamente classificadas em classe 3, das 390 amostras de classe 4, 384 foram corretamente classificadas como 4 e 6 foram classificadas em classe 3. Das 113 amostras de classe 5, 87 foram corretamente classificadas, 26 amostras foram identificadas como classe 4.

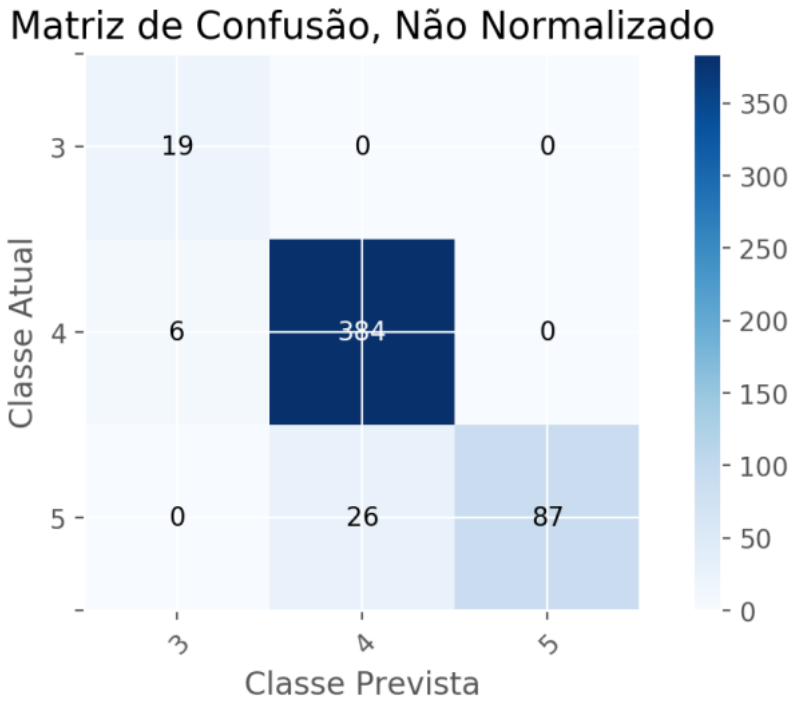

Figura 2 – Matriz de confusão, kNN.

Validação da Precisão

$$Precisão = \frac{\text{TP}}{(TP + FP)} = \frac{19 + 384 + 87}{(19 + 384 + 87) + (6 + 26)} = 0.938$$

3.2.2. Método SVM

O Método SVM, apresentou uma precisão de 92.9% e atribuiu ao novo dado introduzido a classe de previsão, 5 Segundo a matriz de confusão da Figura 10, as 21 amostras de classe 3, foram corretamente classificadas em classe 3, das 378 amostras de classe 4, 368 foram corretamente classificadas como 4, 7 classificadas em classe 3 e 3 foram classificadas na classe 5. Das 123 amostras de classe 5, 96 foram corretamente classificadas, 27 amostras foram identificadas como classe 4.

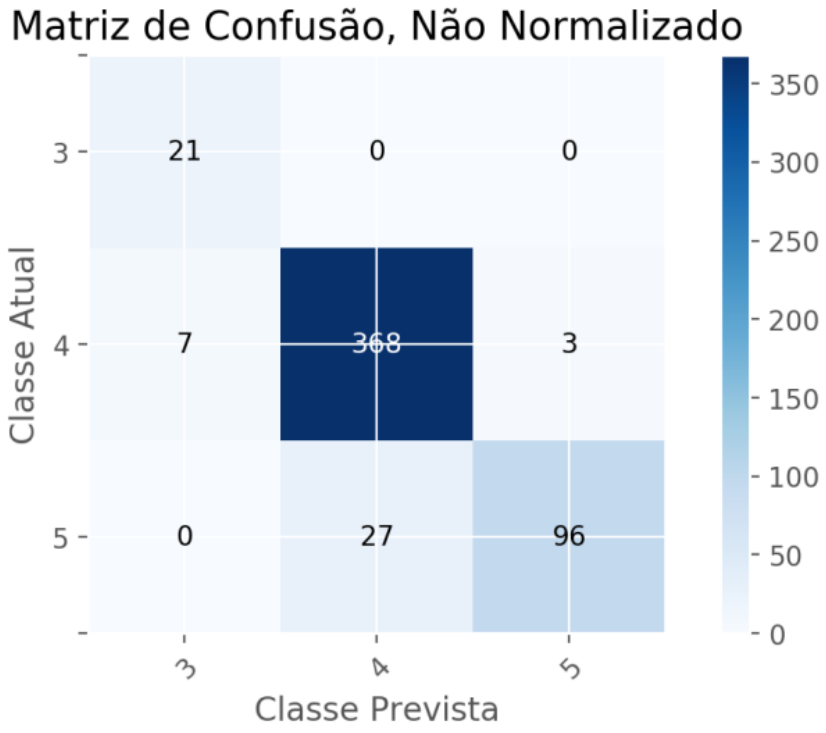

Figura 3 – Matriz de confusão, SVM.

Validação da Precisão

$$Precisão = \frac{\text{TP}}{(TP + FP)} = \frac{21 + 368 + 96}{(21 + 368 + 96) + (7 + 3 + 27)} = 0.929$$

3.2.3. Método RL

O Método Regressão Logística, apresentou uma precisão de 90.8% e atribuiu ao novo dado introduzido a classe de previsão, 5. De acordo com a matriz de confusão da Figura 11, as 18 amostras de classe 3, foram classificadas em

classe 4, das 388 amostras de classe 4, 385 foram corretamente classificadas como 4, 2 foram classificadas em classe 3 e 1 em classe 5. Das 116 amostras de classe 5, 89 foram corretamente classificadas, 27 amostras foram identificadas como classe 4.

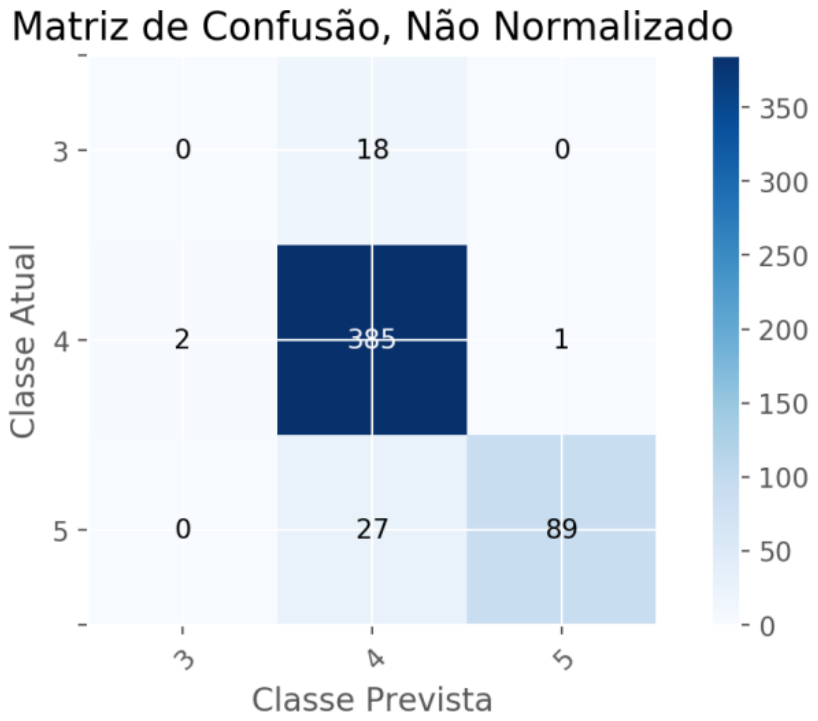

Figura 4 – Matriz de confusão, RL.

Validação da Precisão

$$Precisão = \frac{\mathrm{TP}}{(TP + FP)} = \frac{385 + 89}{(385 + 89) + (18 + 3 + 27)} = 0.908$$

3.2.4. Método NB

O Método probabilístico Naive Bayes, apresentou uma precisão de 94.6% e atribuiu ao novo dado introduzido a classe de previsão, 5. Segundo a matriz de confusão da Figura 12, as 18 amostras de classe 3, foram corretamente classificadas em classe 3, das 392 amostras de classe 4, 389 foram corretamente classificadas como 4 e 3 foram classificadas em classe 3. Das 112 amostras de classe 5, 87 foram corretamente classificadas, 24 amostras foram identificadas como classe 4 e 1 na classe 3.

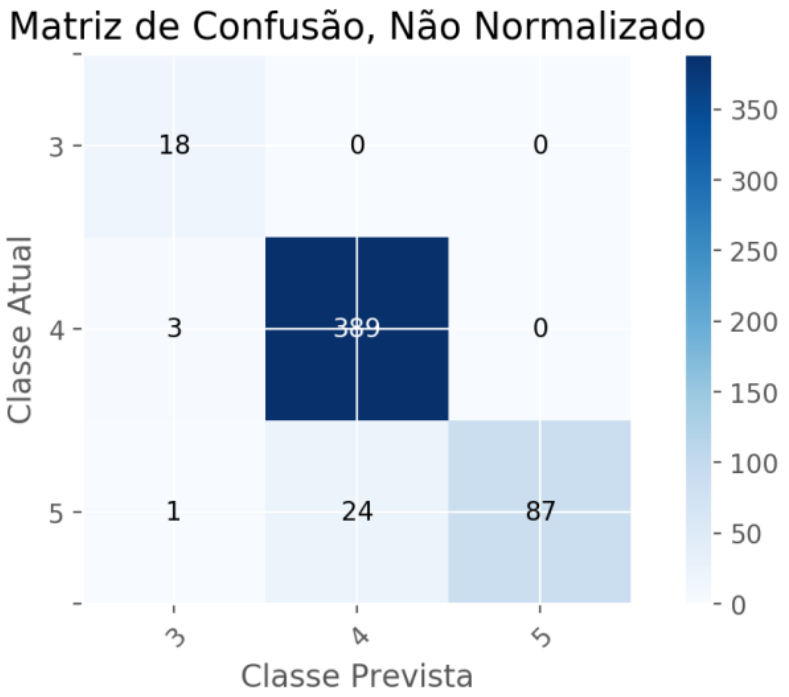

Figura 5 – Matriz de confusão, NB.

Validação da Precisão

$$Precisão = \frac{\text{TP}}{(TP + FP)} = \frac{18 + 389 + 87}{(18 + 389 + 87) + (3 + 1 + 24)} = 0.946$$

3.2.5. Método NB

O Método probabilístico Árvore de Decisão, foi avaliado com uma precisão de 94.4% e atribuiu ao novo dado introduzido a classe de previsão, 3. Segundo a matriz de confusão da Figura 13, as 20 amostras de classe 3, foram corretamente classificadas em classe 3, das 393 amostras de classe 4, 385 foram corretamente classificadas como 4 e 8 foram classificadas em classe 3. Das 109 amostras de classe 5, 88 foram corretamente classificadas, 20 amostras foram identificadas como classe 4 e 1 em classe 3.

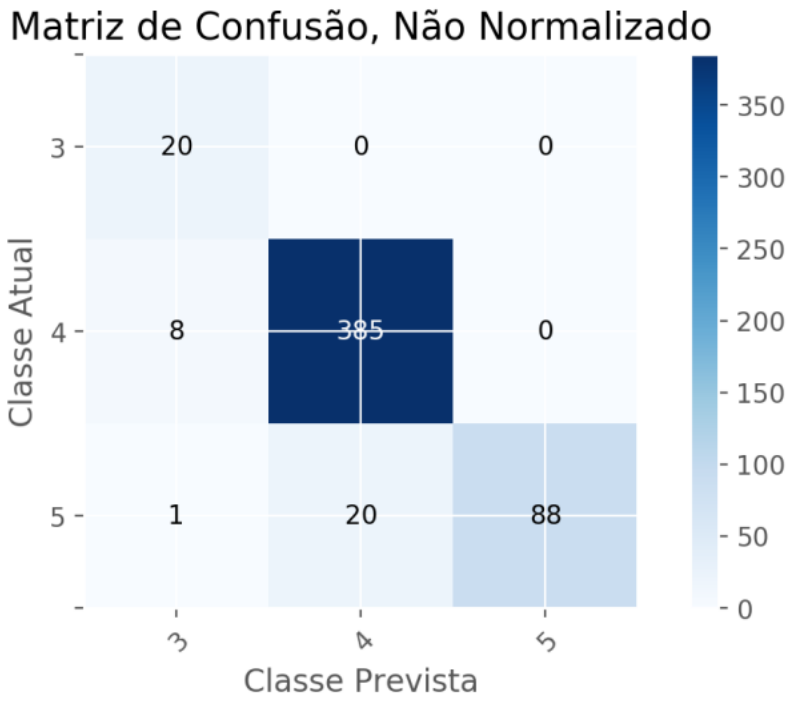

Figura 6 – Matriz de confusão, AD.

Validação da Precisão

$$Precisão = \frac{\text{TP}}{(TP + FP)} = \frac{20 + 385 + 88}{(20 + 385 + 88) + (8 + 1 + 20)} = 0.944$$

3.3. Discussão dos Resultados

Neste capítulo faz-se a comparação dos resultados obtidos dos diferentes métodos de classificação, avaliar o comportamento dos classificadores após introduzir uma nova entrada (valores que correspondem a cada característica da parcela de solo) como também identificar quais as características das parcelas de solo que influenciam mais os resultados obtidos dos classificadores.

Segundo a literatura, o método SVM é o que melhor se aplica tendo em conta a quantidade de dados presente e o facto de ser um estudo que utilizou métodos de classificação para prever classes de aptidão. Após a análise dos resultados e a validação dos mesmos este estudo mostra que o método que prevê uma melhor classe de aptidão é o método NB (*Naive Bayes*).

A comparação dos métodos de classificação permitiu avaliar o método que gerou da melhor forma a classe após introduzir uma nova entrada. De acordo com a tabela 2 podemos analisar que o método RL apresenta uma métrica de precisão inferior comparada com os restantes métodos. Em relação ao método AD,

atribuiu à nova entrada classe 3, contudo apresenta uma métrica superior à da maioria dos métodos, seguindo esta análise os métodos RL e SVM com métricas de precisão inferiores atribuíram à nova entrada, classe 5, a mesma previsão gerada pelo método NB.

Em relação às características que influenciam mais os resultados gerados pelos métodos de classificação, posso afirmar que as características físicas/químicas e biológicas a espessura efetiva como também o risco de inundação são as que influenciam mais os resultados, de modo que, alterando um valor das mesmas o classificador gera outra classe. Para tal, os classificadores que utilizei para comprovar a afirmação acima revelada foi o método de classificação SVM e o método AD.

Os resultados encontrados no presente estudo sugerem que, para uma melhor classificação, previsão e precisão dos métodos, a quantidade de dados recolhidos é bastante importante, quer isto dizer que, como os métodos de previsão utilizam parâmetros de percentagem de dados de teste e treino, quanto mais dados existirem melhor estes métodos classificam as novas entradas introduzidas gerando assim resultados mais fidedignos. Os métodos introduzidos geraram previsões na ordem dos 90 e 95 por cento, o que significa que os mesmos estão bastante otimizados e preveem classes de aptidão para novas entradas de uma forma fidedigna.

3. CONCLUSÕES

Este trabalho vem na sequência de outros projetos anteriormente desenvolvidos com objetivo de implementar, avaliar e validar os diferentes modelos de classificação para prever uma classe de aptidão de solos de modo a otimizar um sistema de rega de precisão com uma contribuição importante para a qual se pretende alargar o regadio da Cova da Beira.

Este estudo é baseado na informação das características dos solos, projeto que foi desenvolvido em 2016 em uma tese de dissertação de mestrado apresentada Escola Superior Agrária do Instituto Politécnico de Castelo Branco [11].

Numa fase inicial o principal desafio foi interpretar e analisar os dados com recurso à linguagem de programação *Python*, posteriormente ultrapassei

diversas dificuldades na implementação dos classificadores bem como as respetivas análises de resultados através de matrizes de confusão.

Na sequência deste trabalho, podem ser desenvolvidos projetos na qual o objetivo é construir uma base de dados digital com a informação das características dos solos e introduzir novos dados, de forma a gerar automaticamente classes de previsão para os mesmos com recurso a modelos de classificação com o intuito de fornecer uma resposta mais eficaz na altura de gerar uma previsão.

Como nota final, foi um trabalho realmente desafiante pelo facto de contribuir ou não na decisão de implementar um regadio a sul do concelho do Fundão, como também pelo facto de abordar este capítulo da Inteligência Artificial que é o *Machine Learning*.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] LIAKOS, K.G., P. BUSATO, D. MOSHOU, S. PEARSON, D. BOCHTIS. 2018. **Machine Learning in Agriculture: A Review**. Sensors, 18, 2674. https://doi.org/10.3390/s18082674.

[2] COPELAND, M. 2016. **A Diferença Entre Inteligência Artificial, Machine Learning e Deep Learning, Dara Sciense Brigade**. Disponível online em: https://medium.com/data-science-brigade/a-diferen%C3%A7a-entre-intelig%C3%AAncia-artificial-machine-learning-e-deep-learning-930b5cc2aa42, consultado em julho de 2022.

[3] STIMPSON, A. J., M. L. CUMMINGS. 2014. **Assessing intervention timing in computer-based education using machine learning algorithms**. IEEE Access 2: 78-87.

[4] ARULKUMARAN, K., M. P. DEISENROTH, M. BRUNDAGE, A. A. BHARATH. 2017. **Deep Reinforcement Learning: A Brief Survey**, in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26-38, , doi: 10.1109/MSP.2017.2743240.

[5] ZHANG, Z. 2016. **Introduction to machine learning: k-nearest neighbors**. Ann Transl Med., 4(11):218. doi: 10.21037/atm.2016.03.37. PMID: 27386492; PMCID: PMC4916348.

[6] CERVANTES, J., F. GARCIA-LAMONT, L. RODRÍGUEZ-MAZAHUA, A. LOPEZ. 2020. **A comprehensive survey on support vector machine classification: Applications, challenges and trends**, Neurocomputing, Volume 408, Pages 189-215, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2019.10.118.

[7] SPERANDEI, S. 2014. **Understanding logistic regression analysis**. Biochem Med (Zagreb). 15;24(1):12-8. doi: 10.11613/BM.2014.003. PMID: 24627710; PMCID: PMC3936971.

[8] MURPHY, K. P. 2006. **Naive bayes classifiers**." University of British Columbia 18.60: 1-8.

[9] SONG, Y. Y., Y. LU. 2015. **Decision tree methods: applications for classification and prediction**. Shanghai Arch Psychiatry. 27(2):130-5.doi: 10.11919/j.issn.1002-0829.215044. PMID: 26120265; PMCID: PMC4466856.

[10] DGADR. 2015. **Sistema de informação do regadio**. Direção Geral de Agricultura e Desenvolvimento Rural. Disponível em: http: //sir.dgadr.pt/expl_centro.

[11] MARQUES, S. G. 2016., **Avaliação da Aptidão dos Solos a Sul do Concelho do Fundão com vista à sua beneficiação pelo regadio, com o uso de ferramentas SIG**. Dissertação para a obtenção de grau mestre na Escola Superior Agrária do Instituto Politécnico de Castelo Branco. Disponível em: https://webcache.googleusercontent.com/search?q=cache:j5rCGodffCsJ:https://repositorio.ipcb.pt/bitstream/10400.11/5333/1/Thesis.pdf+&cd=13&hl=pt-PT&ct=clnk&gl=pt

.