UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA VETERINÁRIA

VALIDATION OF RESISTOME SIGNATURES THROUGH THE APPLICATION OF A MACHINE LEARNING PREDICTION ALGORITHM ON METAGENOMIC DATA

HELENA SOFIA FERNANDES SALGUEIRO

ORIENTADORA:
Doutora Ana Sofia Ribeiro Duarte

TUTOR:
Mestre Telmo Renato Landeiro Raposo Pina Nunes

2023

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA VETERINÁRIA

VALIDATION OF RESISTOME SIGNATURES THROUGH THE APPLICATION OF A MACHINE LEARNING PREDICTION ALGORITHM ON METAGENOMIC DATA

HELENA SOFIA FERNANDES SALGUEIRO

DISSERTAÇÃO DE MESTRADO INTEGRADO EM MEDICINA VETERINÁRIA

JÚRI
PRESIDENTE:
Doutora Maria Manuela Oliveira Castilho Monteiro de Oliveira

VOGAIS:
Doutora Maria Isabel Neto Cunha Fonseca
Doutora Ana Sofia Ribeiro Duarte

ORIENTADORA:
Doutora Ana Sofia Ribeiro Duarte

TUTOR:
Mestre Telmo Renato Landeiro Raposo Pina Nunes

2023

DECLARAÇÃO RELATIVA ÀS CONDIÇÕES DE REPRODUÇÃO DA DISSERTAÇÃO

Nome: Helena Sofia Fernandes Salgueiro

Título da Tese ou Dissertação: Validation of resistome signatures for source-attribution of antimicrobial resistance using machine learning prediction algorithms

Ano de conclusão (indicar o da data da realização das provas públicas): 2022

Designação do curso de Mestrado ou de Doutoramento: Mestrado Integrado em Medicina Veterinária

Área científica em que melhor se enquadra (assinale uma):

☐ Clínica ☐ Produção Animal e Segurança Alimentar

☐ Morfologia e Função ☒ Sanidade Animal

Declaro sobre compromisso de honra que a tese ou dissertação agora entregue corresponde à que foi aprovada pelo júri constituído pela Faculdade de Medicina Veterinária da ULISBOA.

Declaro que concedo à Faculdade de Medicina Veterinária e aos seus agentes uma licença não-exclusiva para arquivar e tornar acessível, nomeadamente através do seu repositório institucional, nas condições abaixo indicadas, a minha tese ou dissertação, no todo ou em parte, em suporte digital.

Declaro que autorizo a Faculdade de Medicina Veterinária a arquivar mais de uma cópia da tese ou dissertação e a, sem alterar o seu conteúdo, converter o documento entregue, para qualquer formato de ficheiro, meio ou suporte, para efeitos de preservação e acesso.

Retenho todos os direitos de autor relativos à tese ou dissertação, e o direito de a usar em trabalhos futuros (como artigos ou livros).

Concordo que a minha tese ou dissertação seja colocada no repositório da Faculdade de Medicina Veterinária com o seguinte estatuto (assinale um):

1. ☒ Disponibilização imediata do conjunto do trabalho para acesso mundial;
2. ☐ Disponibilização do conjunto do trabalho para acesso exclusivo na Faculdade de Medicina Veterinária durante o período de ☐ 6 meses, ☐ 12 meses, sendo que após o tempo assinalado autorizo o acesso mundial*;

* Indique o motivo do embargo (OBRIGATÓRIO)

Nos exemplares das dissertações de mestrado ou teses de doutoramento entregues para a prestação de provas na Universidade e dos quais é obrigatoriamente enviado um exemplar para depósito na Biblioteca da Faculdade de Medicina Veterinária da Universidade de Lisboa deve constar uma das seguintes declarações (incluir apenas uma das três):

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.
2. É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA TESE/TRABALHO (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.) APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.
3. DE ACORDO COM A LEGISLAÇÃO EM VIGOR, (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.) NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA TESE/TRABALHO.

Faculdade de Medicina Veterinária da Universidade de Lisboa, 25 de Maio de 2023

(indicar aqui a data da realização das provas públicas)

Assinatura: 

ACKNOWLEDGEMENTS

It is funny how causality works. There are certain people in my life that have carved the person I am today. They did it with small actions, reactions, with words that engraved in my memory. I didn't even notice it myself. We never notice it when it’s happening. They were the ones who shaped the way I speak, how I write, how I think. In a way, they indirectly shaped this work. They will continue to do so, until we are all specks of dust in the universe. Throughout our lives, we don't express enough times how important people are. We all want to mean something to someone. This is my opportunity to express, to these certain people, how meaningful they are to me.

To my mother. For being the most wonderful, strong human being to have by my side. Life challenged you many times, but you kept going. You kept going, smiling and taking care of others. You never told me how I should act in life, you showed me. You are a nurturer at heart. I thank you with all my heart for showing me unconditional love every day of my life, and for showing me that everything in life can be overcome with kindness and love. You are my everything everywhere all at once.

To my father. For being an example in everything you undertake. Your dedication and hard work has always been something I aspired to for myself, and you are responsible for sowing those aspirations. Thank you for constantly challenging me to be better than myself. For keeping me down to earth, and constantly reminding me that there's still so much I don't know, but that’s ok because you will teach me. Thank you for exploring the world with me. You have a hard shell but a soft heart. Thank you for all the times you let me access it. For all the times you expressed pride in who I am, you have no idea what that meant to me.

To my godparents. Which is to say, to my second parents. Thank you for being home, for supporting me and loving me. Thank you for teaching me that parent love has nothing to do with biology. You raised me as your own. So there will forever be a little bit of you in me.

To my stepfather. For treating me like a daughter. For all the hard work you dedicate, to me and my mother. For being one of the most selfless people I could have come across in this life. I feel grateful for having you in my life.

To my stepmother. For being incredibly patient and always facing life with a cheerful spirit. For really caring about me. Thank you for the joy and lightness you show every day.

To my cousins. As an only child, you were the brothers and sisters I never had. You were role models to me, each in your own way. Thank you for the family gatherings, at the tiny table away from the adults. For sticking together through the stage of life when our clothes were chosen by our parents. I am genuinely proud of you all. Thank you for the warmest Christmases. Thank you for our childhood.

To my grandparents. You are at the origin of everything. At the origin of my genetic predispositions, but also at the origin of my learning, of my curiosity. I remember the first time I understood as a child, from the stories you told me, that older people had very old stories and had lived much longer than I had. You were the first people in the entire world to teach me, without even knowing it, that all people have unique stories and journeys. You nurtured my curiosity about what stories would I have to tell one day, to my grandchildren. I hope you are proud, even if you cannot witness this moment.

To all my middle school friends. For the unique moments we spent in class, which are still subject of conversation at dinners and gatherings. For experiencing together the terrible teenage years. For showing me that a group of fifteen people like ours can stick together for twenty years, counting. Thank you, for all the fun, for all the laughter.

To my girlfriends in particular. I have to thank you for all the beautiful feminist women you have become. You are my therapy and I love you.

To my college friends. For sharing with me all the pain and all the fears of failing. For all the support we have given each other. You were the best thing that happened to me in my college years. To my internship colleagues in particular. For the teamwork, for all the brain storming. Thank you for the friends for life I have now.

And, because who I am today is not limited to family and friends... There are many people I would like to dedicate a paragraph of these acknowledgments to, but there are two in particular that I need to highlight.

To Professor Telmo Nunes. For challenging me to improve my critical thinking and autonomy. For not giving me answers right away. Thank you for being a great teacher. You have no idea of the impact you have on students' lives. Sometimes we just need someone who really cares. I genuinely wish that everyone could have a mentor like you. I have to thank you for opening doors for me and encouraging me to fly.

Finally, to Dr. Ana Sofia Duarte. The person without whom this project would not come to life. Thank you for welcoming me with open arms. For accompanying me on this adventure that was Denmark, which became a huge personal learning experience. Thank you for sharing your personal experience of living abroad. You weren't just a supervisor on my journey. If you had only been a supervisor, you would be the best supervisor I could hope for. But you weren't just a supervisor...

RESUMO

Validação de *resistome-signatures* através da aplicação de um algoritmo de previsão de *machine learning* em dados metagenómicos

Dados metagenómicos têm sido cada vez mais usados em estudos de resistência aos antimicrobianos, mas ainda há uma escassez de métodos precisos e fidedignos para prever a atribuição relativa de genes de resistência a diferentes espécies animais. A disponibilidade de dados de resistência aos antimicrobianos aumentou exponencialmente nos últimos anos, assim como a consciencialização global sobre a ameaça que as resistências representam para a saúde pública, geralmente conhecida como pandemia silenciosa. Isto levou a um aumento no interesse em aplicar métodos de *machine learning* a esses dados.

Neste estudo, sequências *shot-gun* foram usadas a partir de amostras fecais de porcos, frangos, perús e vitelos, recolhidas anteriormente durante estudos nacionais por toda a Europa. Os dados utilizados neste estudo corresponderam a essas amostras e os seus valores *FPKM* associados. Um modelo de *random forest* (RF) foi desenvolvido para prever a atribuição relativa de gene de resistência para essas diferentes espécies. Além disso, uma análise descritiva foi feita para investigar melhor as 15 variáveis mais importantes para o modelo de RF. Uma análise de componentes principais (PCA) e regressão *all-subsets* foram realizadas para identificar genes de resistência específicos de certas espécies. Por fim, esses genes específicos aqui identificados foram comparados com os *resistome-signatures* identificados num estudo anterior.

Os nossos resultados demonstraram que o modelo classificou com sucesso as amostras em classes de espécies correspondentes, com alta precisão e confiabilidade. O modelo teve mais dificuldade em diferenciar porco de vitela, e frango de perú, indicando uma semelhança da composição do resistoma entre cada uma dessas duas espécies. Esta análise validou vários genes como *resistome-signatures* de animais específicos, como *tet(40)* e *sul2* de vitelos, *tet(Q)*, *mef(A)* e *cfxA2* de vitelos e porcos, bla_{TEM}*-126* de frangos, e *tet(A)* de frangos e perús.

Este estudo descreve um método confiável e preciso para a atribuição relativa de genes de resistência a diferentes reservatórios animais usando dados metagenómicos. Estes resultados são essenciais para a vigilância e controlo das resistências aos antimicrobianos em populações animais e humanas.

Palavras-chave: *random forest*; *machine learning*; metagenómica; resistência aos antimicrobianos.

ABSTRACT

Validation of resistome signatures through the application of a machine learning prediction algorithm on metagenomic data

Metagenomic data has been increasingly used in antimicrobial resistance (AMR) studies, but there is still a need for accurate and reliable methods for predicting the relative attribution of AMR determinants to different animal reservoirs. AMR data availability has increased exponentially over the past few years, as has global awareness of the threat that AMR poses to public health, often known as the silent pandemic. This has led to an upsurge in interest in applying machine learning to AMR data.

In this study, shot-gun sequences were used from fecal samples of pigs, broilers, turkeys, and veal calves, previously collected during national cross-sectional studies across Europe. The data used in this study corresponded to these samples and their associated relative abundance of AMR determinants. A random forest (RF) model was developed to investigate the relative attribution of AMR determinants to those different reservoirs. Additionally, a descriptive analysis was made to further investigate the 15 most important variables for the RF model. A principal component analysis (PCA) and all-subsets regression were performed to identify reservoir-specific AMR determinants. Ultimately, the reservoir-specific AMR determinants identified here were compared with the resistome signatures identified in a previous study.

The results demonstrated that the RF model successfully classified resistomes into corresponding reservoir classes, with high accuracy and reliability. The RF model had more difficulty differentiating pig from veal and broiler from turkey, indicating the similarity of resistome composition between each of these two species. The analyses validated several AMR determinants as resistome signatures of specific animal reservoirs, such as *tet(40)* and *sul2* of veal, *tet(Q)*, *mef(A)* and *cfxA2* of veal and pig, bla_{TEM}*-126* of broiler, and *tet(A)* of broiler and turkey.

This study describes a reliable and accurate method for the relative attribution of AMR determinants to different animal reservoirs using metagenomic data. Such results are essential for effective surveillance and control of AMR in animal and human populations.

Keywords: random forest; machine learning; metagenomics; antimicrobial resistance.

RESUMO ALARGADO

Validação de *resistome-signatures* através da aplicação de um algoritmo de previsão de *machine learning* em dados metagenómicos

A resistência aos antimicrobianos é uma grande preocupação global para a saúde humana e animal. Dados metagenómicos têm sido cada vez mais usados em estudos de resistência aos antimicrobianos, mas ainda há uma escassez de métodos precisos e fidedignos para prever a atribuição relativa de genes de resistência a diferentes espécies animais. A disponibilidade de dados de resistência aos antimicrobianos aumentou exponencialmente nos últimos anos, assim como a consciencialização global sobre a ameaça que as resistências representam para a saúde pública, geralmente conhecida como pandemia silenciosa. Isto levou a um aumento no interesse em aplicar métodos de *machine learning* a esses dados.

A presente dissertação nasceu de um crescente interesse por novas abordagens para investigar a ameaça da resistência aos antimicrobianos. Este projeto concentra-se em explorar um método de *machine learning* (*Random Forest*) para prever a fonte mais provável de *clusters* de genes de resistência numa determinada amostra e investigar o grupo de genes de resistência que mais contribuiu para fazer essa previsão, possivelmente revelando *resistome signatures* específicos de cada reservatório. Por fim, os resultados deste trabalho foram comparados com um estudo anterior, com o objetivo de validar as suas principais conclusões.

Random Forest (RF) é um método ensemble, que combina múltiplas árvores de decisão. Uma árvore de decisão é um algoritmo que, usando uma amostra aleatória dos dados, divide consecutivamente as observações dessa amostra ao longo da árvore, desde o nodo raíz até aos nodos terminais, que deverão ser mais homogéneos. No final de cada árvore, é feita uma classificação para cada observação da amostra. Portanto, de forma simplificada, um RF treina múltiplas árvores de decisão, centenas ou milhares, de forma a criar um só modelo mais estável e mais preciso. Várias amostras aleatórias passam por várias árvores de decisão, de forma a treinar o modelo a fazer previsões, sendo que no final há um voto de maioria para obter as classificações finais. Cada observação é classificada com base no que a maioria das árvores votou.

As amostras utilizadas neste estudo foram previamente recolhidas em vários estudos transversais nacionais, no âmbito do projeto europeu EFFORT. Resumidamente, as amostras foram recolhidas em explorações de suínos, frangos, perús e vitelos, de 2014 a 2016, em nove países europeus. Para cada exploração, um total de 25 amostras fecais frescas foram coletadas e agrupadas num único *pool*, de modo a que cada exploração fosse representada por um *pool* de 25 amostras. Seguidamente, as amostras foram sequenciadas e alinhadas com a base de dados ResFinder, resultando em sequências que representam o resistoma

adquirido de cada amostra. No final, os valores de FPKM foram calculados para cada sequência com a finalidade de ajustar as diferenças dos comprimentos dos genes e do número de leituras por amostra, e a abundância de pares de leitura foi agregada correspondendo a 90% de grupos de identidade de genes (*AMR determinants*). Portanto, após limpeza dos dados, os dados utilizados para construir o modelo são constituídos por 389 *AMR determinants* individuais distribuídos por 479 explorações.

Um modelo Random Forest (RF) foi aplicado para classificar a fonte dos resistomas em classes de reservatório correspondentes, com base na sua composição relativa de *AMR determinants*. Primeiramente, os dados foram divididos e foi feito um ajuste dos hiperparâmetros do modelo (duzentas árvores, dez a trinta como o número de variáveis usadas em cada divisão de cada árvore, validação cruzada de dez vezes como o método de reamostragem, impureza de Gini como a regra de divisão). 80% dos dados foram utilizados para treinar o modelo, sendo possível ver qual o número de variáveis que o modelo escolheu utilizar em cada divisão, analisar os valores de precisão e de erro, e obter a lista de variáveis que foram mais importantes para o modelo classificar as amostras, sendo essas as variáveis que vão ser exploradas. Os restantes 20% foram usados para testar o modelo, de modo a fazer previsões com um modelo já treinado, e analisar a sua performance com valores de MCC (*Mathews Correlation Coefficient*).

Após serem feitos vários testes para determinar a confiabilidade do modelo, uma análise descritiva foi feita para investigar as 15 variáveis mais importantes para o modelo RF. Uma análise de componentes principais (PCA) foi feita com o objetivo de perceber como as amostras se agrupam e as variáveis mais responsáveis pela variância dos dados, e uma regressão *all-subsets* foi realizada com o objetivo de analisar de forma mais linear quais as variáveis que mais explicam cada uma das diferentes espécies e identificar genes de resistência específicos dessas espécies. Por fim, foi atribuída uma fonte a cada um dos genes com base nestas análises e os resultados deste trabalho foram comparados aos resultados de um estudo anterior, de forma a validar os diferentes *resistome signatures*.

Os resultados deste estudo demonstraram que o modelo classificou com sucesso as amostras nas espécies correspondentes, com um valor de MCC de 0.97, indicando alta precisão e confiabilidade. O modelo teve mais dificuldade em diferenciar porco de vitelo, e frango de perú, indicando uma semelhança da composição do resistoma entre cada uma dessas duas espécies. Uma possível razão para esta dificuldade e para o rácio de amostras erradas de perú e vitelo, é que as amostras de perú e de vitelo representam uma minoria da base de dados, o que significa que o modelo tem menos informação para aprender sobre estas duas espécies, possivelmente explicando o porquê de as confundir com frango e porco, respetivamente.

Os testes realizados para testar a confiabilidade do modelo demostraram, de uma forma geral, que o modelo é preciso e generalizável e que a importância das variáveis atribuída pelo modelo não foi atribuída de forma aleatória, isto é, os quinze clusters de genes são realmente essenciais ao modelo para fazer previsões.

Após comparação dos quinze clusters de genes identificados neste estudo com os *resistome-signatures* identificados no estudo anterior, observou-se que oito clusters foram identificados em ambos os estudos e atribuídos de forma geral às mesmas espécies, validando a sua caracterização como resistome-signatures.

Em suma, as análises realizadas neste estudo validaram vários clusters como *resistome-signatures* de animais específicos. Em particular, *tet(40)* e *sul2* de vitelos, *tet(Q)*, *mef(A)* e *cfxA2* de vitelos e porcos, bla_{TEM}*-126* de frangos, e *tet(A)* de frangos e perús. O cluster *cfxA6* foi considerado específico de porco neste estudo, mas considerado específico de humano no estudo anterior. Esta diferença deve-se ao facto de o estudo anterior ter incorporado amostras de humanos que contactassem diretamente com animais em explorações e matadouros, e poderá indicar que humanos apresentam um resistoma mais semelhante ao de porco, em comparação com outras espécies.

Este estudo descreve um método confiável e preciso para a atribuição relativa de genes de resistência a diferentes reservatórios animais usando dados metagenómicos. Sem dúvida, este estudo valida vários *AMR determinants* como *resistome-signatures* de fontes animais específicas, contribuindo para informar o desenvolvimento de intervenções direcionadas para prevenir a resistência aos antimicrobianos. Estas descobertas destacam a importância da análise complementar na investigação de marcadores genéticos de resistência e os desafios de diferenciar entre certas fontes.

Estes resultados são essenciais para a vigilância e controlo das resistências aos antimicrobianos em populações animais e humanas, demonstrando o enorme potencial da aplicação de modelos de machine learning no âmbito das resistências aos antimicrobianos.

Palavras-chave: *random forest*; *machine learning*; metagenómica; resistência aos antimicrobianos.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

adjr2	Adjusted R-square
AI	Artificial intelligence
AMR	Antimicrobial resistance
AMU	Antimicrobial use
DTU	Technical University of Denmark
FAO	Food and Agriculture Organization of the United Nations
FPKM	Fragments Per Kilobase of reference and Million Reads
GAP	Global Action Plan
GDP	Gross domestic product
ICU	Intensive care unit
MCC	Matthews correlation coefficient
MDA	Mean decrease in accuracy
MDI	Mean decrease in impurity
ML	Machine learning
mtry	Number of variables used at each split
NGS	Next-generation sequencing
OIE	World Organization for Animal Health
OOB	Out-of-bag error
PCA	Principal component analysis
RF	Random forest
UNEP	United Nations Environment Programme
WGS	Whole genome sequencing
WHO	World Health Organization

1. CHAPTER I – INTERNSHIP REPORT AND PROJECT INTRODUCTION

1.1. Internship report

The curricular internship that gave rise to this project and dissertation was carried out from January 17th to May 13th, 2022, in Kongens Lyngby, Denmark. This internship was supervised by Dr. Ana Sofia Ribeiro Duarte, senior researcher at the Technical University of Denmark (DTU), the institution that hosted me during this period.

Since the curricular internship required some background knowledge on R language and data science, another internship was previously carried out from September 1st to December 10th, 2021, at the Faculty of Veterinary Medicine - University of Lisbon, under the guidance of Dr. Telmo Nunes. This served as a training period where, with other colleagues, I learned how to code in R Studio and the basis of data handling, competences that were essential for the development of this project. We worked with several datasets, mainly focusing on statistical data analysis, data transformation and data visualization, within the scope of veterinary epidemiology. We also helped to analyze the progress of the COVID-19 situation in our faculty and the best way to communicate and present it. This was an important time to consolidate the importance of teamwork, autonomy, and self-learning.

While in Denmark, the first couple of weeks of the curricular internship were spent meeting the team and work accommodations and debating about the project that would turn out to be this dissertation. After investing some time analyzing the provided dataset, I proceeded to navigating deeper into the world of coding. With the additional help of Dr. Derya Aytan-Aktug, I was able to understand machine learning methods, how we apply them in the field of epidemiology, and also to create and test a random forest model.

During this time in DTU, I also enrolled in a course on “Surveillance and Epidemiology of Infectious Diseases”, in which every Monday we worked on different exercises regarding epidemiology of zoonoses, burden of disease, surveillance systems of infectious diseases and antimicrobial resistance, among other. This was a complementary tool that helped to improve my knowledge in the field of epidemiology and public health.

A team meeting was held every week, where I was able to follow the progress of the other team members' work, as well as share my own progress. Every week I also had meetings with my supervisor, where I was able to express my difficulties, discuss the following steps and refine timelines. The internship ended on May 13th and on May 16th I gave a final presentation to the team, in order to share an overview of my internship and project activities, as well as to receive constructive insights from other professionals in this area.

In order to make this experience possible, financial aid was provided by the Erasmus+ programme.

1.2. Project introduction

Every time we use antimicrobials in people, animals and plants, microorganisms have the opportunity to develop the ability to tolerate the treatments by becoming resistant, making the drugs less effective over time. Antimicrobial resistance (AMR) is a major global concern to human and animal health. It has implications for food safety, food security and the economic wellbeing of millions of farming households.

The present dissertation was born out of a growing interest for new approaches to investigate the threat of antimicrobial resistance. This project focuses on exploring a machine learning method to predict the most likely source of resistance in a given sample (the resistance reservoir), and to investigate the group of resistance genes that contributed the most to make that prediction, thus possibly revealing reservoir specific genomic resistance markers. Ultimately, the results of this work were compared to a previous study (Duarte et al. 2021), with the aim of validating its main findings.

The first step in this project consisted in learning about antimicrobial resistance, the importance of understanding its epidemiological mechanisms and forms of prevention, and finally learning about machine learning methods under the scope of epidemiology. This knowledge was gathered through a literature review, corresponding to chapter II of this dissertation.

In chapter III, the modelling approach is described in detail, followed by chapters IV and V, where the results are presented and discussed, respectively.

2. CHAPTER II – LITERATURE REVIEW

2.1. Antimicrobial resistance: a threat to global health

2.1.1. Brief history of antimicrobials and antimicrobial resistance

Perhaps for the general public the history of antimicrobials begins with the discovery of penicillin in 1928. Indeed, this discovery was one of the greatest milestones in medicine, but it was not the beginning of antimicrobials. It could be said that the history of antimicrobials does not have a beginning. In ancient Egypt, moldy bread and honey were used on infected wounds (Pećanac et al. 2013), and Traditional Chinese Medicine suggested the use of calcarea and musty mixtures for infected burns (Kopp et al. 2003). Moreover, investigation of human skeletal remains from ancient Sudanese Nubia (350-550 CE) revealed traces of tetracycline in their bones, evidencing this population possible intention of obtaining nutritional and pharmacological effects from food fermentation (Basset et al. 1980; Nelson et al. 2010). These studies allow us to understand that, since ancient times, human beings have been curious about exploring antimicrobial properties long before we discovered the first bacterium.

During the 19th century, the effects of these molds and fungi were further explored and in 1897, thirty years ahead of Fleming, Ernest Duchesne discovered that *Penicillium glaucum* had the ability to inhibit *Escherichia coli* (Durand et al. 2018). Nonetheless, the era of antimicrobials starts having a solid foundation with Paul Ehrlich and his search for a "magic bullet" that selectively targeted only the pathogenic microbes of a host. It was the beginning of the 1910s, when he introduced Arsphenamine, also known as Salvarsan or compound 606, for the treatment of syphilis, endemic and incurable at the time. Although it was considered chemotherapy, this was the first modern purified antimicrobial molecule (Williams 2009).

Little breakthroughs like these made us one step further for the discovery of antimicrobials as we know them, and this takes us to the discovery of penicillin by (Fleming 1929). This serendipitous finding happened by accident, when Alexander Fleming found forgotten colonies of *Staphylococcus aureus* and noticed that *Penicillium notatum* was inhibiting their growth (Bennett and Chung 2001). This fungus' molecule was isolated and industrialized as penicillin only years after, by Chain and Florey (1944). However, before the industrialization of penicillin, the discovery and industrialization of the sulfa drugs also took place. Sulphonamide, synthesized years earlier by Josef Klarer and Fritz Mietzsch, became massively produced in 1936 (Aminov 2010). Sulphonamide and its derivatives were widely used by soldiers during World War II along with penicillin (Durand et al. 2018).

By this time in history, the benefits that antimicrobials could bring to humanity had become quite clear. They were pioneers in the control of diseases considered leading causes of morbidity and mortality throughout all human existence.

The success of these drugs led to their increasing and indiscriminate use across the world, and with that a new problem began to emerge. It became evident that more and more bacteria and other microbes became resistant to them. Comparing the timeline of antibiotic introduction with the appearance of antibiotic resistance, illustrated in Table 1, it is noticeable how rapidly resistance mechanisms have followed our mass consumption of antimicrobials. Davies and Davies (2010) contemplate the distribution of resistance in microbial populations as "the result of many years of unremitting selection pressure from human applications of antibiotics, via underuse, overuse, and misuse". The magnitude of this selection pressure triggered by humans is a key factor for the surge of generations of resistant microbes, which have biological advantages over susceptible populations (Oz et al. 2014; Gillings et al. 2017).

Table 1 - Timeline comparison of antimicrobial introduction and identification of resistant microbes (CDC 2019)**.**

Antimicrobial approved/released	Year released	Resistant Germ Identified	Year identified
Penicillin	1941	Penicillin-resistant *Staphylococcus aureus*	1942
		Penicillin-resistant *Streptococcus pneumoniae*	1967
		Penicillinase-producing *Neisseria gonorrhoeae*	1976
Vancomycin	1958	Plamsid-mediated vancomycin-resistant *Enterococcus faecium*	1988
		Vancomycin-resistant *Staphylococcus aureus*	2002
Methicillin	1960	Methicillin-resistant *Staphylococcus aureus*	1960
Extended-spectrum cephalosporins	1980	Extended-spectrum beta-lactamase-producing *Escherichia coli*	1983
Azithromycin	1980	Azithromycin-resistant *Neisseria gonorrhoeae*	2011
Imipenem	1985	Klebsiella pneumoniae carbapenemase (KPC)- producing *Klebsiella pneumoniae*	1996
Fluconazole	1990 (FDA approved)	Fluconazole-resistant *Candida*	1988

Despite becoming increasingly frequent after the widespread use of antimicrobials, the phenomenon of antimicrobial resistance was in fact not novel. The isolation of bacteria from water and ice samples with more than 2000 years showed resistance to ampicillin (Dancer et al. 1997). Metagenomic analysis of isolates from permafrost sediments with more than 30000 years old reported resistance genes against vancomycin, tetracycline, and others (D'Costa et al. 2011). Studies on the phylogeny of β-lactamases determined that these enzymes originated more than 2 billion years ago and were capable of remaining in plasmids for millions of years (Hall and Barlow 2004; Garau et al. 2005). According to research, the phenomenon of antimicrobial resistance precedes the antimicrobial era, so much so that its occurrence accompanies the existence of microbes themselves. It appears that soil-dwelling microbes, mostly non-pathogenic bacteria, have evolved mechanisms of interaction between them. When exposed to chemicals and threats from other microbes, they are able to develop defense mechanisms, such as the production of bioactive molecules, which later evolve into specific resistance elements (Morrison and Zembower 2020). D'Costa et al. (2006) demonstrated this phenomenon by isolating spore-forming bacteria from soil, and subjecting them to exposure to various antibiotics, both natural and synthetic. All strains proved to be multi-resistant, even to antibiotics currently still used as gold-standard treatment and antibiotics only recently industrialized. This is one of many studies that pinpoint the occurrence of antimicrobial resistance in environmental bacteria from the soil and their potential role in the transfer of resistance to pathogenic bacteria, considering monitoring of resistance in the environment a promising strategy to predict and prevent future emergence of resistant strains.

The mechanisms by which the transfer of resistance between microbes occurs are many. Some microorganisms can be intrinsically resistant to some antimicrobials, as Cox and Wright (2013) portrays. In the process of multiplying, microbes copy their genetic material millions of times, where mutations frequently happen, slightly altering the genomes. More often than not, mutations have little to no consequence in the fitness of the microbes. But, sometimes, a mutation can be the key for resisting an antimicrobial. The mechanisms by which they become resistant include preventing access of the antimicrobial into the organism, removing the antimicrobial using pumps in the cell wall, destroying or inactivating the antimicrobial with enzymes, changing the antimicrobial target so that it cannot latch on to it, or developing new processes that avoid the microorganism being targeted by the antimicrobial (Munita and Arias 2016; Varela et al. 2021). Although this is a natural process for microbes to develop intrinsic resistance, the epidemiology of resistance is accelerated due to repeated exposure to antimicrobials, which increase the probability of resistant bacteria to develop, adapt and survive. Resistant bacteria can multiply, pass the resistance determinants to new generations and can also share their genetic material with other bacteria. This is called horizontal gene transfer and most microbes acquire resistance genes through these

mechanisms. It can occur through transduction, when the transfer of genes occurs via bacteriophages; conjugation, when genes are transferred using plasmids or transposons, upon direct contact between microbes; or transformation, when microbes pick up resistance genes that were released in the environment by other live or dead microbes (Rossi et al. 2014; Pulingam et al. 2022).

Antimicrobial resistance has been around for as long as microbes have had to resist each other, but it is safe to say that the mass use of these wonder drugs for human benefit was a turning point for the epidemiology of resistance. Between the 1940s and 1960s, the golden period of antibiotic discoveries led to a global wave of optimism that bacterial infections could be eradicated. However, during the 1970s, the discovery of new drug classes decreased significantly, while the resistance problematic continued to expand. This forced the scientific community to investigate enhancements upon the existing collection of drugs, initiating the fabrication of synthetic drugs and combining the effects of different preparations (Chopra et al. 2002). While this approach continues to be successful in providing effective treatments, the past has provided us with a window into the future. As we try to keep up with resistance, by modifying existing drugs and finding new ones, microbes will evolve to resist our prevailing antimicrobials. This arms race between humans and microbes is one that will persist for a long time to come.

2.1.2. Epidemiology of antimicrobial resistance in veterinary practice

Microbes co-exist in communities called microbiota, which are found not only in people and animals, but also in all environmental elements around us. Research on the behavior of these communities, especially how microbes interact with each other and with the environment in order to evolve, can help us understand how resistance develops.

The fact that antimicrobials are so present in our daily lives can mislead the public perception to the false idea that antimicrobials are mainly used for purposes related to human diseases. In fact, of all the applications of antibiotics, it is estimated that less than 50% correspond to therapeutic use in humans (Davies and Davies 2010). Ever-increasing amounts of antimicrobial agents have been used, among other practices, in prophylaxis and growth promotion in animal production, treatment and prophylaxis in household pets and aquaculture, as pest control in agriculture, as biocides in some household cleaning products and even in sanitizers and disinfectants, a use that was magnified during the COVID-19 pandemic period (Cabello et al. 2016; Xiong W, Sun Y, et al. 2018; Marteinson et al. 2022). All these examples have an impactful role in the propagation of resistance, with animal production being one of the most consequential. According to Palma et al. (2020), during the last decades the companion animal population has significantly increased in developed countries, and so has the demand on food production due to population growth. On one hand, this has led in many

instances to improvements in animal health care. On the other hand, it has resulted in an increased use of antimicrobials in animals and in an increased contact between animals and humans, contributing to a high risk of transmission of zoonotic diseases and AMR bacteria.

The role of food industry and animal production as a leading contributor for the epidemiology of antimicrobial resistance has been well documented over the years (Verraes et al. 2013; Haulisah et al. 2021; Saraiva et al. 2022). The use of antibiotics in our small domestic animals is considered relatively similar to the use in humans, as most antibiotics are used to treat infections or for prophylactic use, such as post-surgery. This is not the case for food-producing animals, to which antimicrobials are mostly administered in subtherapeutic doses to large groups in the feedlot and drinking water, as a way of preventing certain diseases or to promote their growth. Although the use of antimicrobials as growth promotors has been banned in Europe and other countries many years ago, China, the largest producer and consumer of antimicrobials, was using 50% of this consumption to promote animal growth up until 2013 (Rahman et al. 2022).

The use of antimicrobials implies their discharge through several paths such as hospital and clinical waste, animal husbandry, manufacturing industry, among others. The major concern about using antimicrobials in animals is that up to 90% of these drugs are eliminated in their urine and feces, being released into the environment as manure with other residues from farms, contaminating water, soil and agriculture products (Samreen et al. 2021). The massive amounts of antimicrobials circulating on farms and, consequently, on the environment poses a particular threat when it comes to resistance. As Palma et al. (2020) clarify, exposing bacteria to constant sublethal doses causes a perfect selective pressure that encourages the adaptation and evolution of resistant pathogenic microorganisms.

The network of interactions between microbes and antimicrobials happening everywhere all at once leads to resistance reaching humans through countless pathways. For instance, hospital environments can be saturated with resistant pathogens, as there is a large available genetic pool of resistance (Dancer 2014). Other transmission routes occur through the food chain. Resistance can spread through direct or indirect contact between animals or animal products and humans, being farmers, slaughterhouse workers and food processors the most exposed and affected, as shown by Van Gompel et al. (2020). Farmed animals can result in contaminated retail food products. Despite the withdrawal period when using antimicrobials, some residues can persist in meat, meat products, fish and dairy products, leading to the development of resistant bacteria in the human digestive tract (Manyi-Loh et al. 2018). Some animal food products can also carry resistant bacteria. For example, antibiotic-resistant strains of *Salmonella* have been found in broiler, beef, pork and eggs (Nair et al. 2018). Resistant *Campylobacter spp*. is a concern in many livestock species, and other concerning zoonotic resistant bacteria include *Staphylococcus aureus*, *Listeria monocytogenes* (Kozajda et al.

2019; Abebe et al. 2020) and extended-spectrum beta-lactamase producing *E. coli* (DANMAP 2021).

A numerous diversity of microorganisms can equally be found in water, soil and air. Regardless of their pathogenicity, these microorganisms are excellent sources of resistance genes and can act as reservoirs of the environmental resistome. A “resistome” is defined by Wright (2007) as the collection of all resistance genes present in a given environmental sample and plays a potential important role in transference of resistance genes to pathogenic microbes. The presence of resistant microorganisms in water and soil is partially a consequence of farm discharges and use of manure in agriculture soil (Larsson and Flach 2022). Manure is also used as fertilizer in integrated fish farming, increasing the levels of resistant pathogens in aquaculture environments (Petersen et al. 2002), and in agriculture soil, contaminating fruits and vegetables for human consumption (Uyttendaele et al. 2015; Zhou et al. 2020). Contaminated water from farms and wastewater-treatment plants can reach streams, rivers and other surface waters, possibly exposing humans through swimming to pathogenic bacteria, such as β-lactamase-producing *E. coli* (Schijven et al. 2015; Zhu et al. 2019). Moreover, antibiotic-resistant bacteria have been shown to spread through barn floor, animal transportation vehicles and portable equipment, potentially ending in animal carcasses during slaughter (Dorr et al. 2009).

The epidemiology of antimicrobial resistance helps us acknowledge this is a planet of microbes much more than it is a planet of humans. Microorganisms circulate through the ecosystem through complex pathways, coexisting and interacting with every environmental element along the way.

2.1.3. How antimicrobial resistance threatens global health

At a patient level, the direct consequences of antimicrobial resistance are easily perceivable. Infections by resistant bacteria can double the chance of severe illness and triple the probability of death, when compared to those caused by non-resistant bacteria (Cecchini et al. 2015). According to Murray et al. (2022), around five million deaths were associated with antimicrobial-resistant bacteria in 2019 alone, one million of which were directly caused by AMR. Despite the difficulty to get accurate predictions when it comes to the burden of resistance, some studies report that by the year of 2050 antimicrobial-resistant infections can cause the death of ten million people, if nothing is done to prevent them (Chokshi et al. 2019). We can already witness this trend today. Treatment for tuberculosis and malaria are becoming less efficient, and 60% of HIV infections in Sub-Saharan Africa were already resistant to anti-HIV medicine in 2017, perhaps hindering the goal of putting an end to AIDS by 2030 (O’Neill 2016; WHO and Fund 2017).

The direct outcomes of resistant microbes regarding human health are reflected in a cascade of problems in the healthcare systems. If treating resistant infections is ineffective, patients will need longer hospitalizations, more isolation beds and intensive care units (ICUs). This comes at a high cost, as treatments will become more expensive both for patients and hospitals. Some studies suggest that AMR could cost from $300 billion to $1 trillion worldwide, by 2050 (World Bank 2017; Chokshi et al. 2019). There will also be a burden through secondary effects. Some procedures, like surgeries, require the use of antimicrobials to decrease the risk of post-surgery infection. If the effectiveness of these prophylaxis is compromised, certain medical procedures cannot be performed successfully. This means that diseases which are curable today with surgery, transplants or dialysis, may become incurable (Santoro-Lopes and Gouvêa 2014; Naylor et al. 2018). Likewise, AMR is also going to have an impact on the outcome of diseases like cancer. Chemotherapy weakens the immune system, making cancer patients vulnerable to different infections. If that vulnerability cannot be protected with antimicrobials, chemotherapy can actually become more dangerous and less effective (Dadgostar 2019).

At a social and economic level, the consequences of AMR are many but can sometimes be overlooked. Projections conducted by Taylor et al. (2014) and World Bank (2017) showed that, if trends continue, annual gross domestic product (GDP) could decrease by 1% to 4% globally by 2050, translating into trillions of dollars. Multidrug-resistant tuberculosis alone could cost the world $16.7 trillion by 2050 (World Bank 2017). These studies also revealed that AMR influences labor through the loss of productivity caused by sickness and premature death. Underprivileged populations are the most dependent on labor income, which will be reduced due to the prevalence of infectious diseases (Alsan et al. 2015). This, associated with higher expenses on treatments, will only increase the rate of extreme poverty and low-income countries will suffer the most compared to the rest of the world. AMR can thus highlight the gap between developed and developing countries, further aggravating inequity worldwide.

The impacts on livestock can also be significant. Increased morbidity and mortality rates would cause a decrease in animal production and trade, leading to price inflation for protein sources such as meat, milk and eggs (Hillock et al. 2022). This is alarming as some studies show that the demand for protein sources has increased in recent years (Tiseo et al. 2020). The projections by the World Bank (2017) estimate that, if nothing is done to counteract these trends, animal production could decline by up to 11% by 2050, a substantial loss that will incite a decline in income and intensify the economic situation. Once more, an economic factor affecting people in extreme poverty enhancing wealth disparities and inequality. Some authors even consider the long-term impacts of AMR on the global economy to be more severe than a global financial recession (World Bank 2017).

2.1.4. Importance of prevention

It is a privilege to live in a time after the discovery of antimicrobials. It has truly been a significant milestone for humanity, one that we should not take for granted.

Antimicrobial resistance is such a complex phenomenon that its reduction and prevention require a coordinated multisectoral approach. There has been some progress in the past years. Thanks to advocacy and sustained research conducted by the Tripartite, constituted by the World Health Organization (WHO), the Food and Agriculture Organization (FAO) and the World Organization for Animal Health (OIE), a consensus on the approach strategies is summarized in the Global Action Plan on AMR (WHO 2015). In 2016, during the seventy-first session of the United Nations General Assembly, political leaders issued a high-level political declaration committing to implement the Global Action Plan (GAP) at a global, regional and national level. This action plan is routinely monitored and re-evaluated, to assess necessary changes of strategies and to ensure all stakeholders maintain their responsibilities and leading experts keep guiding their countries on better national AMR policies. The Global Action Plan outlines five major points, detailed below.

The first objective stands to improve awareness and understanding of antimicrobial resistance to the general public. This includes effective communication among human and animal health audiences, and also incorporating AMR in professional education and training. The public communication and public media also play an important role in helping to prioritize AMR in government's agenda (WHO 2015).

Secondly, strengthening the knowledge and evidence base through surveillance and research. Developing an AMR surveillance system for healthcare, animal husbandry and agriculture is the foundation of a strategic prevention (Aarestrup et al. 2008). A good example is Denmark, which developed a highly integrated and systematic program, (DANMAP 2021), that takes into account antimicrobial use and resistance in every step of the food chain. DANMAP was the first national surveillance system to be introduced by a country and since then has been reproduced by other countries, such as RESAPATH in France, CIPARS in Canada and NARMS in the United States. DANMAP employs unique methods of integrating data, being shared with the international public and conceiving outcomes for action through cross-sector collaboration between scientists and authorities (Simjee et al. 2018).

Another goal of the Global Action Plan is reducing the incidence of infections through effective sanitation, hygiene and infection prevention measures. We currently have healthier guidelines over treatment of industrial, civil and farm waste, and our global concern about animal welfare and hygiene has led to better living conditions for animals and structural changes in all food industry sectors (Collignon 2013). These measures are particularly important in healthcare facilities and farms, and include promoting food and water hygiene,

implementing better practices in terms of animal health, testing and reporting hospital-acquired infections, and promoting vaccination of food animals (WHO 2015).

The fourth ambition is to optimize the use of antimicrobial agents in human and animal health. There have been some advancements in the past, such as the prohibition of growth-promoters in Europe, and other antibiotic restrictions (Cecchini et al. 2015). However, antimicrobials are often prescribed empirically for human use, in some instances even sold over-the-counter (World Bank 2017). Moreover, antimicrobials are still most times overused in food industry and agriculture. This is why the Global Action Plan strives for the evidence-based prescribing through the enhancement of rapid and low-cost diagnostic tools, the execution of stewardship programs, and the implementation of superior regulation over distribution and use of antimicrobials, adjusted to the reality of each country (WHO 2015).

The ultimate target of the GAP translates in ensuring the economic case for sustainable investment that takes account all countries' needs. The aim is to: encourage more assessments on the socioeconomic burden of antimicrobial resistance; invest in the development of new medicines and vaccines; engagement in research collaborations between developed and developing countries (WHO 2015). These actions require a whole society approach with all sectors and disciplines working together, i.e., it calls for a One Health approach incorporating human health, animal health, and environmental factors.

In 2022, the United Nations Environment Programme (UNEP) joined the former Tripartite (FAO, OIE and WHO) as an equal partner. The Quadripartite collaboration signed a new Memorandum of Understanding (MoU), which provides legal framework for the four international agencies to further their shared approaches on AMR at the human-animal-ecosystem interface (FAO et al. 2022). This is another example of efforts made by the scientific community to strengthen the already existent cooperation to sustainably balance and optimize global health.

2.1.5. Future surveillance of antimicrobial resistance using metagenomics

Collecting information systematically from laboratories, hospitals and surveys, that allow us to detect, monitor and characterize infectious diseases has always been part of public health. However, there is a need to better understand and monitor outbreaks of AMR in order to achieve effective public health strategies. Recent advances in microbial genomics, such as Next-Generation Sequencing (NGS) methods, revolutionized the way DNA sequencing could be applied to food safety and public health, and have paved the way for the application of genomic methods such as Whole Genome Sequencing (WGS), Shotgun Metagenomics and Single Cell Sequencing in AMR surveillance (Boolchandani et al. 2019; Trotter et al. 2019). Genomic surveillance is already in use for outbreak investigation (Hendriksen R, Munk P et al.

2019). What was once an academic pursuit, has become an increasingly powerful tool in helping public health surveillance.

Whole Genome Sequencing, one of the most used methods for surveillance, is the ordering of an organism's entire set of nucleotides within their DNA (Ng and Kirkness 2010). The advantages of WGS can be summarized by its unprecedented level of detail, as resistant bacteria can even be traced by specific allele profiles, rather than phenotypic resistance patterns to a selection of drug classes (Zhao et al. 2016). It has been used in AMR surveillance to rapidly detect emerging threats (Gerner-Smidt et al. 2019; Zhang et al. 2020; Tagliani et al. 2021). However, WGS-based AMR surveillance focuses on specific individual bacteria, resulting in a rather limited microbial spectrum which may not capture all the relevant resistance genes circulating in a microbial community (Hendriksen R, Munk P et al. 2019). The constituents of the antimicrobial resistome have contributions from the microbial community as a whole, rather than from an individual source microorganism (Abreu et al. 2021). Shotgun Metagenomics, in particular, has facilitated the analysis of antimicrobial resistance in complex microbial communities, opposing to culture-based techniques, such as phenotypic testing or WGS.

Metagenomics is a sequencing-based analysis of genomes, defined by the direct extraction of the total DNA from a microbial community contained within an environmental sample (Riesenfeld et al. 2004). Metagenomic techniques are able to identify thousands of resistance genes in a single sample, without the need to pre-select which genes to look for, unravelling thus the so called resistome (Boolchandani et al. 2019). The obtained community taxonomic profile can provide massive amounts of information on genomic assembly, gene prediction, species diversity, among others (Cowan et al. 2015). If novel resistance genes are identified in the future, we can go back to this library and re-analyze it for the presence of genes we had not identified yet (Hendriksen R, Munk P et al. 2019). Moreover, since metagenomics provides culture-independent insights of microbial communities, it has emerged as an attractive monitoring tool, that allows to capture the full spectrum of microbial diversity (Duarte et al. 2020). Research has shown that metagenomics is superior to conventional methods for quantifying resistance genes in sludge or manured soil (Huang et al. 2014; Fang et al. 2015), for AMR surveillance of swine herds (Xiao et al. 2016; Munk et al. 2017), for comparison of resistome diversity across livestock in Europe and for epidemiological analysis of AMR risk factors (Munk et al. 2018; Van Gompel et al. 2019). This approach was also demonstrated valuable in global AMR surveillance using urban sewage (Petersen et al. 2015; Hendriksen R, Bortolaia V et al. 2019; Aarestrup and Woolhouse 2020). Furthermore, Duarte et al. (2021) demonstrated that metagenomic data can be used to predict the relative attribution of AMR determinants observed in human resistomes to different livestock reservoirs and to identify reservoir resistome markers, by using random forest models. Machine learning

algorithms have also been applied to metagenomic data, with a major purpose of producing real-time reliable predictions (Gupta et al. 2019).

As NGS technologies become more affordable, it becomes increasingly realistic to use metagenomics to investigate the potential role of different environments in the ecology of resistance. Hence, metagenomics supports the prospect of a One Health surveillance of the microbial ecosystem where data and predictions are generated and reported in a culture-independent manner, analogous to a real-time "weather map" of infectious diseases and resistance.

2.2. Machine learning: a tool for AMR prediction and surveillance

2.2.1. Historical context of machine learning and its application in AMR

Robotics as a concept has history that extends for thousands of years, to the ancient Egyptian culture and Greek mythology, long before the term *robot* was coined by Karel Capek (Gasparetto 2016). This concept of machines being able to execute simple and repetitive labor for humans remained present until after the industrial revolution, when scientists, mathematicians and philosophers started hypothesizing that human thought and reasoning could be mechanized (Boole 1854; Frege 1879). Based on these theories, Russel and Whitehead (1910) published the masterpiece "Principia Mathematica", which sparked an interest in the formalization of mathematical reasoning and popularized the subject of symbolic logic. By the early 20th century, in parallel with the discovery and industrialization of penicillin, artificially intelligent robots were being portrayed in fiction books and movies, such as the humanoid robot in *Metropolis*,1927, or Tin man from the *Wizard of Oz*, 1939 (Behnke 2008). When Turing (1950) published his famous essay "Computing Machinery and Intelligence", this concept that was part of our imagination for thousands of years became tangible. In his essay, Alan Turing discussed how we could build intelligent machines and introduced the Turing test, a method for testing their intelligence. This thesis turned out to be a foundation for artificial intelligence (AI) and machine learning (ML).

Nonetheless, the term "artificial intelligence" or "AI" was only coined in 1956 in the *Dartmouth Summer Research Project on Artificial Intelligence*, organized by Marvin Minsky and John McCarthy, where the new AI research discipline was founded (Lu 2019). They defined artificial intelligence as "the science and engineering of making intelligent, thinking machines" (McCarthy et al. 2006). Currently, the Dartmouth conference is widely contemplated as the birth of AI as we know it, catalyzing the following years of revolutionary research.

By that time in history, AI research focused on creating instruction-based systems, which used predefined rules to make decisions. However, these systems did not have the ability to learn from data and adapt to different circumstances, as the lack of computational power of that time proved to be a major obstacle (McCorduck and Cfe 2004).

It was only during the 80s that AI had a resurgence, with machine learning at its core (Smith et al. 2006). Going back to 1950, Turing proposed that instead of writing a computer program that imitates humans, we should write a program that imitates the way humans learn (Turing 1950). Machine learning is precisely the implementation of Turing's proposal. It is a subfield of artificial intelligence that focuses on understanding and building analytic models which can learn from past experience, with little or no human intervention (Bi et al. 2019). Thus, instead of building a program to translate from English to Portuguese, we should write a program that is able to learn by observing examples of translations in both languages, much like how a child acquires language (Herlau et al. 2021). Machines have become progressively able to learn based on observed data, whether from spreadsheets, images, sensors, or direct human instructions. Hopfield (1982) and Rumelhart et al. (1985) further popularized deep learning techniques, which allowed computers to accomplish tasks and generalize the learned ability to future similar situations. Machine learning opened the door for the accomplishment of groundbreaking ambitions of artificial intelligence.

Today, machine learning is interwoven into people's daily lives. We are surrounded by AI systems that make our daily tasks easier. A person can use a virtual assistant on their phone to order food from an online application, which was recommended by advertisements specifically tailored for that person, paying with a credit card which is automatically checked for fraud (Herlau et al. 2021). This is just the tip of the iceberg. We live in the era of Big Data and the current massive technological power gave our computers the storage capability and processing speed required to learn from those amounts of inputs and expand machine learning to every scientific field (Duan et al. 2019).

In the comprehensive field of global health, machine learning has broadly intervened in the detection of lesions and medical diagnosis (Andrade et al. 2010; Jaeger et al. 2018; Moyo et al. 2018), morbidity/mortality risk assessment and prediction of certain diseases' progression (Phakhounthong et al. 2018; Johnston et al. 2019), infectious diseases outbreaks prediction and surveillance (Kabaria et al. 2016; Haddawy et al. 2018; Clemente et al. 2019), and improvement of health policy and planning (Rosas et al. 2013; Yousefi et al. 2018).

Interest in applying machine learning to antimicrobial resistance data has intensified over the past years, reflecting the exponential increase of AMR data available and the increasing global awareness of the public health threat posed by AMR – also called the silent pandemic. These methods can use clinical information of patients and antimicrobial susceptibility data (Oonsivilai et al. 2018; Martínez-Agüero et al. 2019), or genomic data (Moradigaravand et al. 2018; Nguyen et al. 2019).

As Dick (2019) ponders, the history of artificial intelligence goes beyond the mechanical attempts to imitate the static notion of human intelligence, as it also reflects the change in how we perceive intelligence itself. With the exponential rise of artificial intelligence, some ethical

concerns have been legitimized (Harlow 2018). Despite this, the scientific community agrees on the important benefits machine learning can generate, such as liberating humanity from the burden of repetitive work, whether this is accomplished by robots or machine learning tools for data processing, in order to help us expand our capabilities.

2.2.2 Supervised machine learning

The classification of machine learning can be broadly summarized in whether the program's learning is supervised, unsupervised, or by reinforcement. This review will focus on supervised machine learning, specifically random forests, but for the sake of completeness, unsupervised and reinforcement learning are going to be briefly addressed.

Starting with reinforcement, this is a type of learning where an AI model interacts with an environment that rewards its sequence of decisions with a negative or positive result, allowing the model to store that new information and make reinforced responses or decisions in the future (François-Lavet et al. 2018). Depending on how it is employed, this approach can be comparable to supervised ones (Herlau et al. 2021). It has been increasingly used in game engines, image processing, natural language processing, among others (Sharma and Kaushik 2017; Kaiser et al. 2019; Furuta et al. 2020). A good example of reinforcement learning are the intelligent robots Atlas and Spot, from Boston Dynamics (Guizzo 2019).

In unsupervised learning, given a set of observations (dataset), the algorithm independently infers hidden patterns of the features (input data), without having access to the real outcomes or classification (output data) (Herlau et al. 2021). It discovers the underlying structure of the dataset, by finding important features that help to group those data according to similarities, and represent results in a compressed format (Hastie et al. 2009). This type of learning not only can handle large amounts of cluttered and uncategorized data, requiring little data pre-processing, but also identify previously undetected patterns (Hastie et al. 2009). The purpose of unsupervised learning can be: clustering the data, used for example in costumer segmentation or grouping bacteria based on their genome (Sun et al. 2018; Sinaga and Yang 2020); dimensionality reduction, used to select important features of a dataset (Reddy et al. 2020); finding association rules, exploited for example to discover which conditions can imply high risk for future diseases (Santoso 2021); anomaly detection, used for fraud surveillance or prediction of mechanical maintenance (Hilal et al. 2022).

Finally, with supervised machine learning the opposite happens. While unsupervised models train on unlabeled data and are able to act independently without supervision or feedback, the process of supervised learning requires the collecting and preparing of a labeled dataset. The algorithm knows from the start the category of each observation in the training set. This training data works as a supervisor that teaches the model to infer correlations between the features and categories (Herlau et al. 2021). After learning the patterns of the

input data, the model can make predictions about new, unseen data (Cunningham et al. 2008). Supervised learning is the most common type of machine learning and can be broadly classified into two problem categories, regression or classification.

Regression modeling is a method useful to predict continuous values, such as salary or weight. It determines the strength of the relationship between independent variables (inputs or predictors) and one dependent variable (output or response variable) (Berk 2008). The main goal is to find a mathematical function that represents that relationship in a way that the model is able to predict on new data (Berk 2008). A simple example is the linear regression, where there is an assumption that the relationship between input and output variables is linear, and the model finds the line that best defines this relationship (Kumari et al. 2021). Another example is the logistic regression, a variation of linear regression applied in binary classification problems. Logistic regression uses a logistic function to model the probability of an output variable being one of two classes, demonstrating the parameters that maximize the likelihood of the observed data (Nasteski 2017). Some practical uses of regression modeling are weather forecasting, stock price trends, or housing price predictions (Sharma et al. 2018; Pérez-Rave et al. 2019; Mahabub et al. 2020).

Classification is a type of supervised learning used to predict categorical values. Given one or more inputs, a classification model will attempt to draw patterns from the observed values and predict the outcome for each observation (Sen et al. 2020). The outcome can either be binary or multi-class. Some examples of classifiers are, among others, logistic regressions, decision trees and random forests. Note that logistic regression and random forests can be considered either a regression or a classification type, depending if the output variable is defined as a continuous value or a categorical variable, correspondingly (Kumari et al. 2021). Classifiers are used in a wide variety of applications, such as image classification, speech recognition of virtual assistants, e-mail spam filtering, or medical diagnosis based on x-rays or MRI scans (Ghosh et al. 2016; Kaur and Sharma 2019; Li et al. 2019; Raza et al. 2021). The following section will delve deeper into how random forests work.

2.2.3. Random forest as a predictive algorithm

A random forest is an ensemble learning method first introduced by Breiman (2001) which combines multiple decision trees with randomness to create a more stable and accurate model. Compared to single models, an ensemble method trains multiple models and combines their outputs to improve its predictive performance. This can be done through either training the same algorithm with different versions of the dataset to minimize overfitting (e.g., bagging or boosting) or training different algorithms in the same dataset relying on different models' strengths without the need to pre-select an algorithm (e.g., Bayesian model averaging or Super Learner) (Dietterich 2000). Bagging, also known as bootstrap aggregation, is an ensemble

method also introduced by Breiman (1996) and has three steps: 1) bootstrapping, where random samples of observations from a training dataset are selected with replacement; 2) multiple training, when the bootstrap samples are trained independently; 3) aggregation, through majority voting, as the class with the majority of votes among the different models is accepted. This method aims to reduce the variance of a single classifier by combining the results of multiple classifiers trained on different random subsets of the data (Farid et al. 2011). Random forest is an extension of Breiman's bagging idea, applied to decision trees. To understand random forest, it is first necessary to understand decision trees.

A decision tree is an algorithm with a tree-like structure, as represented in Figure 1. Although it can be used for both regression and classification problems, this explanation will focus on classification. Decision trees use a bootstrap sample of the original data, and binary splits recursively partition the tree, pushing the samples from a parent node to its two daughter nodes, so that homogeneity among samples increases from parent nodes to daughter nodes (Kotsiantis 2013). The algorithm splits the data until no further splits are possible, either because the node reached a pure homogeneity (all of one class) or there are no more variables upon which to split (Goldstein et al. 2010). The first node is called the root node and the terminal nodes are called leaf nodes. If the predictor variable is continuous, the split of each nonterminal node is determined by a split point (see example in Figure 1). The algorithm uses a random number of variables (also called features) in each split to increase accuracy and randomization (Cutler et al. 2012).

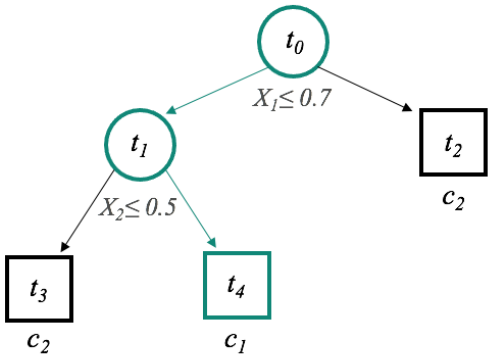

Figure 1 – A decision tree built for a binary classification problem.

The decision tree above has five nodes (t_0, t_1, t_2, t_3, t_4). Node t_0 is the root node and corresponds to the whole input space, or predictor space, X. The split of the root is based on the question "Is $X_1 \leq 0.7$?" which divides X_{t0} into two subsets X_{t1} and X_{t2}. The left node corresponds to the positive answer and the right node to the negative answer. Similarly, t_1 is labeled with the split $X_2 \leq 0.5$ which further divides X_{t1} into two disjoint subsets X_{t3} and X_{t4}, respectively corresponding to the sets of input data that have values ≤ 0.5 and > 0.5. Terminal nodes, or leaf nodes, t_2, t_3 and t_4 are represented by squares and are labeled with an output class, c_1 or c_2. (Louppe 2014)

Additionally to binary classification, decision trees offer the capability of predicting multiple classes. Thus, instead of building an independent binary model for each class, we can build a single multi-class algorithm, predicting all outputs or classes at once (Jha et al. 2019). According to Louppe (2014), the advantage of multi-class decision trees is that existing correlations between output classes are taken into account, whereas individual models do not exploit such dependencies, possibly affecting the accuracy of the model.

A random forest has several properties that makes it advantageous for classification problems. Single decision trees are prone to overfitting, i.e., they can become too complex and memorize the noise in the training data, resulting in poor performance on new, unseen data (Schonlau and Zou 2020). By combining multiple decision trees, a random forest can not only reduce the variance of the predictions, resulting in a more stable and robust predictor, but also reduce the overfitting problem, improving the overall performance of the model (Biau et al. 2008). It can also handle barriers such as missing values and outliers, large number of categorical or numerical input features, multi-class problems, or imbalanced data (Schonlau and Zou 2020). Additionally, random forest has the ability to model non-linear relationships between input and output variables.

In practical terms, when a random forest algorithm is applied to a data set, whatever the programming language, the data set is firstly randomly sorted and then split into a training set and a testing set. Usually, the training set corresponds to 70-80% of the original data which the model uses to learn the variables' patterns, while the testing set corresponds to the other 20-30% of the data which is only used after evaluating the training performance of the model, as a way of testing the algorithm with unseen observations. The random sorting previously mentioned ensures that both the training set and the testing set contain observations belonging to every class, as long as the data is not heavily class imbalanced. The next step is tuning the hyperparameters that control how the model is fit, which involves specifically instructing the model regarding the number of trees it should generate, the depth of each tree (that is, how many nodes each tree should have before it stops) and the number of variables it should investigate in each split and the criteria used to split each node (Schonlau and Zou 2020). The tuning step can also include defining a control resampling method for the training process. One of the most common methods is the k-fold cross-validation, represented in Figure 2. As Ljumovic and Klar (2015) explain, this method divides the training set in k folds or groups, leaving one of the folds as a hold-out set. It fits the model in k-1 folds and then uses the hold-out fold as testing. This is a way for the model to test itself during the training process of each decision tree.

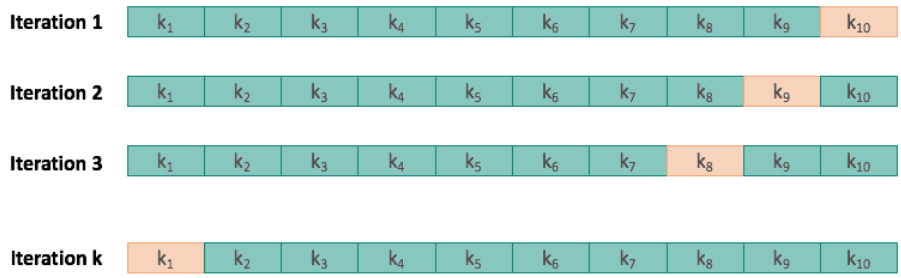

Figure 2 – Representation of the k-fold cross-validation method.

The training data is divided into 10 folds. It uses one of the folds as a hold-out, and only trains using the other 9 folds. After training, it uses the hold-out set to test predictions. This process is repeated k times, each time using a different fold as the hold-out set (Ljumovic and Klar 2015).

After data preparation and model tuning, the training process begins. Random forest can reach a collection of hundreds or thousands of decision trees, with each decision tree using a different bootstrap sample of the training data, as represented in Figure 3. When this collection of trees is generated, a final vote is cast based on the classification of every tree, a process called majority voting (Cutler et al. 2012). Prediction for each observation is thus based on the proportion of votes given to each class across all trees, in the form of relative probabilities, from which the model produces a final classification based on the most likely class, also called the crisp class (Biau et al. 2008).

One of the most important advantages of random forests is its interpretability. It is more easily interpretable than other complex models, as it is possible to extract information about accuracy values, variable importance and relationship between input and output variables. The accuracy of the model is calculated based on the number of observations correctly classified and also based on error estimates, such as the out-of-bag (OOB) error. When each bootstrap sample is selected from the training data, the observations that are left out are called out-of-bag (OOB) sample, which are extremely useful to estimate a generalization error and variable importance (Cutler et al. 2012). Each OOB sample is passed down the tree to produce an estimate prediction error for the sample (Goldstein et al. 2010). It is important to distinguish the OOB error from the k-fold cross validation when evaluating a random forest. Although both mechanisms are identical (both use a set to train the model, and a hold-out to test it), there are some key differences. Not only do they assume different size of learning samples, but also they inform on different probabilities (Hastie et al. 2001). The k-fold cross validation resampling method informs about the probability of the forest giving the correct classification, while the OOB error informs about the probability of the majority vote of forest trees being the correct vote.

Figure 3 – Simplified structure of a random forest.

For training a random forest, N bootstrapped samples, or in-bag samples, are selected from the original dataset. The samples left out of the in-bag sample are called the out-of-bag (OOB) samples. Each bootstrapped sample is then used to grow a classification tree. Then, a small number of randomly selected K predictors are selected in each split. These two steps are then repeated until C such trees are grown, and a classification is obtained by aggregating the prediction of the C trees. (Ahmad et al. 2018)

Of great interest to genetic epidemiologists is the ability of random forest to identify important variables. Variable importance is a measure of how important each variable is for the model's performance and has been receiving increased attention in genomic epidemiology. Since many datasets have hundreds or thousands of predictor variables, it is often a necessary part of prediction model development to select the important variables and reduce the noise in the data. The random forest algorithm has built-in variable importance measures for the training process, such as mean decrease in impurity (MDI) and mean decrease in accuracy (MDA) (Chen and Ishwaran 2012). The MDI calculates, for each variable and for each tree, the decrease in impurity resulting from the split caused by that variable in a given node, taking

into account the number of samples in that node (Loecher 2020). The measure of MDI is then averaged over all trees and normalized (Loecher 2020). The MDA will randomize the rows of each variable in the training set and calculate the decrease in the average predictive accuracy (Louppe 2014). The more it decreases, the more important the predictor variable is. The variables with the highest importance scores are considered the most important variables for the model, worthy of further investigation (Goldstein et al. 2010).

Most machine learning techniques, like logistic regression, support vector machines or random forests, share the characteristic of searching over a transformed version of the feature space in order to find a suitable solution to the problem. Each of these methods has utility in finding an underlying structure in genomic data. Nonetheless, the focus of the current study is the application of a random forest model, not only because it is a straightforward algorithm to apply and interpret, but mainly because of its competence to identify important variables of interest from large datasets. Unsurprisingly, there has been a steady increasing use of random forest in genomic literature since it was first introduced in 2001.

3. CHAPTER III – METHODOLOGY

3.1. Project preparation and main objectives

3.1.1. Sample collection and DNA sequencing

This study did not include sample collection or DNA sequencing. Instead, the samples used in this study were previously collected during several national cross-sectional studies, within the scope of the European project EFFORT. It included farm selection, sampling, pooling of samples, DNA extraction, DNA sequencing and bioinformatics analysis. Although these processes have been already detailed in previous studies (Munk et al. 2018; Duarte et al. 2021), for the sake of completeness this subheading will succinctly address them.

Briefly, the samples used in this study were collected from pig, broiler, turkey and veal farms, from 2014 to 2016. Pig and broiler farms were sampled in nine European countries (anonymized with letters A-I, as agreed by the project consortium), and turkey and veal farms were sampled in three of those nine countries. The total number of pig, broiler, veal and turkey farms sampled in each country is detailed in Table 2. For each farm, a total of 25 fresh fecal samples from pen-floor were collected and pooled in a single pool, so that each farm was represented in a pool of 25 samples.

From each pooled sample, DNA was extracted using a modified QIAamp fast DNA Stool Mini Kit protocol. These DNA samples were shotgun sequenced on the Illumina HiSeq3000 platform, for pigs and broilers, and on the Illumina NovaSeq 6000 platform, for turkeys and veal calves. The bioinformatics analysis of the metagenomic raw reads was similar to the one described by Munk et al. (2018), using a more recent version of the ResFinder database, which is a manually curated database of around 3000 acquired AMR genes (assessed on 21st September 2018). The DNA reads were analyzed with the MGmapper tool and then aligned to the ResFinder database described by Zankari et al. (2012), so that the resulting mapped reads represent the acquired resistome in each sample.

Fragments Per Kilobase of reference and Million reads (FPKM) values were computed for each ResFinder reference sequence in order to adjust for differences in both gene lengths and number of reads in each sample, and the abundance of read pairs was aggregated, corresponding to 90% gene identity clusters (referred to as AMR determinants). The resulting clusters were then manually examined and named accordingly to their representative gene.

In short, the data used in this study is the same raw mapping count data and their associated FPKM values used by Duarte et al. (2021), containing originally 389 individual AMR determinants, among 499 farms sampled.

Table 2 – Number of samples collected in each country (A-I).

Species	A	B	C	D	E	F	G	H	I	Total
Pig	20	20	20	40	20	20	20	20	21	201
Veal	-	21	-	-	20	20	-	-	-	61
Broiler	20	18	20	20	20	20	20	20	19	177
Turkey	-	21	-	-	20	-	-	19	-	60
Total	40	80	40	60	80	60	40	59	40	**499**

3.1.2. Main objectives and study design

This study aims to develop a random forest model to predict the relative attribution of AMR determinants to different animal reservoirs and thus investigate the possible existence of AMR markers of specific reservoirs, with the future perspective of attributing AMR determinants observed in human resistomes to specific transmission sources, i.e., a specific reservoir. Ultimately, the results of this study were compared with those of a study carried out in 2021 by Duarte et al. (2021), in which 16 AMR determinants were designated as resistome signatures, with the purpose of validating those previous findings.

To achieve these goals, firstly the data was pre-processed and split into a train set and a test set. Then, a random forest was developed and tuned, so that it could be fit using the training set. The testing set was used to make predictions, after which the performance of the model was measured, and the important variables were evaluated. The final model was then tested in various ways to validate its consistency. After ensuring the model's reliability, the model's 15 most important determinants were identified. A descriptive analysis of these AMR determinants was performed, and finally the results were compared with those of the previous study conducted by Duarte et al. (2021). The full study design is represented in Figure 4.

All analyses conducted in this study were performed in R Studio v.1.3.1093. The random forest algorithm was performed using the R package *caret* v.6.0-93 (Kuhn 2008) (see further list of R packages used in Appendix I).

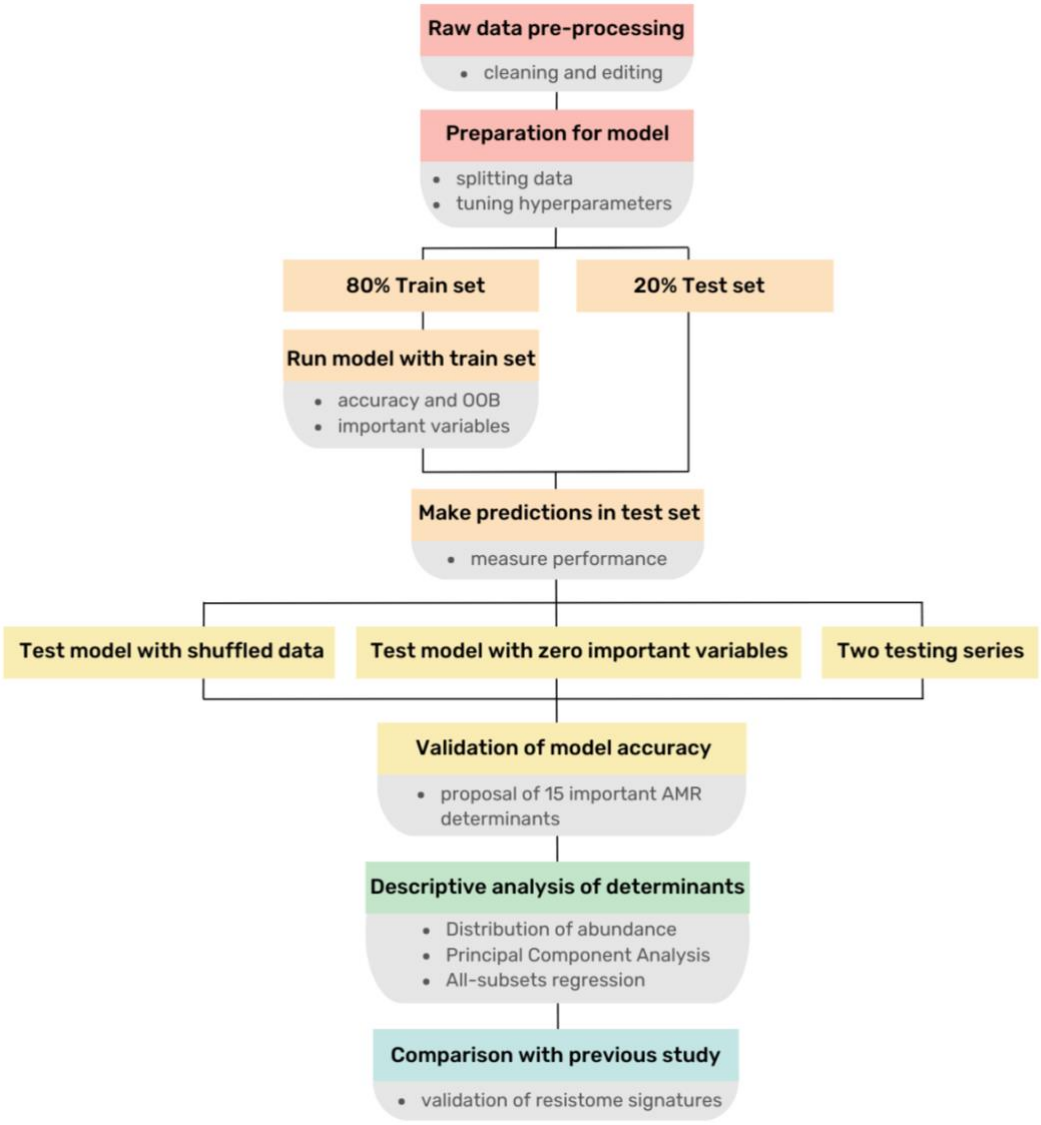

Figure 4 – Study design (original).

3.2. Developing and testing a machine learning algorithm

3.2.1. Data pre-processing

For a random forest model to be applied, it is necessary that the data is specifically formatted. This means having a data frame with the response variable (species) in the first or last column and the predictor variables (AMR determinants) in the remaining columns. This step prepares the data for the forthcoming analyses and embroils data cleansing and editing. Data cleansing involved removing corrupt entries, duplicates, and missing values, while data editing involved transposing the data and configuring all samples and species' names to be correctly formatted.

After all the pre-processing, the data consisted of 389 AMR determinants (variables) with values of abundance present in 479 samples (observations), distributed among 4 different species (181 pig samples, 177 broiler samples, 60 turkey samples, and 61 veal samples).

The AMR determinants are named with the representative genes for the 90% homology cluster.

3.2.2. Development of a random forest model

A Random Forest model was applied in order to classify resistomes into their corresponding reservoir classes, based on their relative composition of AMR determinants. RF is a tree-based supervised machine learning algorithm that is highly adaptive and is able to account for correlation and interaction among explanatory variables (Breiman 2001). The underlying mechanism of random forest is explained in detail in Chapter II, heading “2.2.3. Random forest as a predictive algorithm”.

Firstly, the data was split into 80% training and 20% testing, so that the training set included 384 observations and the testing set included 95 observations. After this, a tuning of the model’s hyperparameters was made, which consisted of (a) establishing the default 200 as the number of decision trees to be generated (*ntree*), (b) selecting 10-fold cross validation repeated 5 times as the resampling method, (c) selecting Gini impurity as the rule to split each node, which is the default for classification, and (d) defining 10 to 30 as the interval of *mtry* values (number of variables used at each tree split), in order to include with some margin the square root of the number of AMR determinants (rule of thumb), from which the model selected the *mtry* value that provided the best performance.

The model was trained using the function *train* of R package *caret* v6.0-93 (Kuhn 2008). This function takes in several arguments, including the training data and all the tuning parameters stated previously. After training for around 30 minutes runtime, the algorithm provided the characteristics of the final model, which *mtry* was best and which samples were most difficult to classify, along with performance measures such as Accuracy and Out-Of-Bag (OOB) error. The built-in accuracy metric of the *train* function is calculated using k-fold cross-validation. The model is trained on k-1 folds and tested on the remaining fold, and this process is repeated k times, with each fold used as the test set exactly once. The accuracy is then calculated as the average proportion of correctly classified instances over all folds. The built-in OOB error of RF is also calculated while the model is training (details in Chapter II, heading “2.2.3. Random forest as a predictive algorithm”).

The final model was fit to the 20% testing set to make predictions, using the *predict* function of *caret*. Prediction for each observation is thus based on the proportion of votes given to each class across all trees, in the form of relative probabilities, from which the model produces a final classification based on the most likely class - the “crisp class”. A confusion matrix was composed based on crisp-class classifications, and prediction performance was assessed using Accuracy and Matthews Correlation Coefficient (MCC). The MCC is a statistical metric used for model evaluation, calculated based on the occurrence of true

positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (Chicco and Jurman 2020). This metric ranges from -1 to 1, where 1 indicates perfect agreement between prediction and observation, -1 indicates total disagreement, and 0 is expected for a prediction no better than random. The MCC was used because it has been shown to be more reliable than other performance metrics, such as the RF accuracy values (Ballabio et al. 2018).

3.2.3. Assessing important variables

After evaluating model performance, the 15 most important variables for classification were checked and plotted. These variables corresponded to AMR determinants. RF importance measures help to understand the significance of each variable in the model (Kuhn and Johnson 2013). For this task, the function *varImp* of R package *caret* v6.0-93 was used (Kuhn 2008).

For each decision tree in the RF model, the reduction in impurity (Gini impurity) is calculated for each variable, every time they split a node. The Gini impurity is a measure of the randomness in the split of the decision tree (Yuan et al. 2021). The *varImp* function calculates the importance score of each variable by averaging the reduction in impurity caused by that variable over all decision trees in the RF model. The more a variable contributes to the reduction in impurity, the higher its importance score. The scores are then scaled such that they range from 0 to 100, to make it easier to compare importance scores of different variables.

3.2.4. Testing the RF model

Before delving into the 15 most important AMR determinants, the model was tested in different ways, to investigate the reliability of its results.

The first approach involved shuffling the sample categories so that the outputs were randomly swapped. In this way, a pig sample could be categorized as turkey or broiler, for example. This made the patterns of abundance no longer related to species. The purpose of this test was to confuse the model and see if its accuracy decreased.

The second approach was performed with correctly categorized samples but using only 15 of the least important variables for the model. There were 127 AMR determinants that the model considered zero important, probably because they have zero or near zero variance of abundance, and 15 of those determinants were randomly selected to test the model. This is another way in which it is possible to check the model is indeed learning from the variable patterns.

A third approach was taken, which comprised two tests series. The first test series included running the model multiple times using only the 15 most important variables for the model, adding one variable at a time. The second test series included the same, but using 15 random variables, which did not include any of the 15 important variables. In short, the model was applied multiple times, with AMR determinants being added one at a time, for each test

series. The MCC values for each test series were plotted. This approach made it possible to evaluate the impact that each individual determinant has in the performance of the model when it is added as a variable and to compare the difference in MCC values between versions of the model using the 15 most important determinants and versions using other 15 randomly selected determinants.

3.3. Descriptive analysis

Descriptive analysis, or descriptive statistics, encompasses methods that allow to describe and visualize the distribution of values from a variable among different observations in a dataset and to aggregate and summarize data in an insightful manner. It generates accessible insights of the data that otherwise would be non-interpretable, by making perceptible the most important information according to the aim of the study.

A descriptive analysis was made to further investigate the most important AMR determinants identified by the RF model. This analysis included (1) investigating the antimicrobial classes to which the 15 AMR determinants belonged, (2) plotting the distribution of abundance of those determinants among the four different species, (3) making a Principal Component Analysis (PCA) to evaluate sample clustering and determinants' influence on data variance, and (4) making an all-subsets regression for evaluating the linear relationship between AMR determinants and each reservoir. The major purpose for this part of the study was to validate the model's outputs as reservoir-specific AMR determinants.

3.3.1. Evaluation of sample clustering and variable influence on data variance

A PCA was performed to evaluate the variable influence on data clustering. PCA is a multivariate algorithm, i.e., it deals with the results of multiple variables for every single observation, which is often used to analyze genomic data because it allows a dimensionality reduction while maintaining most of the variance of the dataset. Dimensionality reduction consists of reducing the number of variables in a dataset without losing significant information, simplifying the data while preserving its essential characteristics. It accomplishes this by projecting the data into a new coordinate system of lower-dimensional subspace (Ringnér 2008). The observations of possibly correlated variables are transformed into clusters of new uncorrelated variables, called principal components (Ringnér 2008). Principal components are generated by deconstructing the original data into eigenvectors and eigenvalues, where the eigenvector is the direction with the highest variance and the eigenvalue is the number that translates the amount of variance in its corresponding eigenvector (Kherif and Latypova 2019). The first principal component refers to the projection of the data to the first axis, or dimension, explaining the highest variance in the data. The second principal component captures the highest variance in the data not captured by the first component. Each subsequent principal

component explains the highest variance not explained by the preceding principal components until all data is decomposed (Kherif and Latypova 2019). Usually, the first two principal components explain most of the variance with minimal loss of information.

Since RF is a complex non-linear method, this study aimed to compare its results to a linear method like the PCA, which can deal with correlated predictors (multicollinearity). A first PCA was made with just the 15 most important AMR determinants as variables, after which a second PCA was made using all data. For this, the function *PCA* from *FactoMineR* v.2.7 package was used (Lê et al. 2008). Each PCA was presented with a plot for the sample clusters in the first two dimensions (the scores plot) and a plot for the variables' position in the same dimensions (the loadings plot). The purpose of this step of the study was to evaluate how samples cluster according to reservoirs and which AMR determinants influence those clusters the most, as well as compare the PCA with the RF's results to investigate if there was any AMR determinant highly influencing the data variance which was not considered essential for the RF model.

3.3.2. Regression analysis for linear evaluation of important variables

A regression analysis consists of a statistical technique widely used to model the linear relationship between a dependent variable and one or more independent variables (Fox 2008). In this study, an all-subsets regression was performed to validate the important variables from the RF model associating important variables with specific reservoirs.

All-subsets regression involves fitting all possible combinations of independent variables to the data and selecting the best formula (or model) to predict the dependent variable based on certain parameters (Miller 2002). All-subsets regression presents several advantages when compared with other regression methods (Wang and Chen 2016), such as providing a comprehensive overview of the relationship between the dependent variable and the independent variables, identifying the best formula among all possible formulas (ensuring the highest goodness-of-fit measure), handling big data with many independent variables, and being easily interpretable.

For each reservoir, an all-subsets regression was made using the *regsubsets* function from the R package *leaps* v.3.1. (Lumley 2003). Accessing the 15 most important AMR determinants for the RF model, the function *regsubsets* searched for the 2 best formulas, using an increasing number of AMR determinants as predictors, to predict pig, veal, turkey, and broiler, individually. The best 2 formulas for each number of predictors were selected based on the adjusted R-squared (adjr2) value. The adjr2 consists of the variance proportion of the dependent variable that can be explained by the independent variables, adjusted to the number of predictor variables in the regression (Miles 2005). This value ranges from 0 to 1, where 0 indicates that the dependent variable cannot be explained by the independent

variables and 1 indicates that the dependent variable can be perfectly explained by the dependent variables. The all-subsets regression resulted in four graphs, illustrating the best formulas for the different subsets of variables in order to predict each reservoir.

3.4. Comparison with previous study

Ultimately, to each RF important AMR determinant was assigned one or two specific reservoirs. The criterion for assigning a specific reservoir to each important AMR determinant prioritized the all-subsets regression, since it finds linear relationships for each species. For an AMR determinant to be considered reservoir-specific, it would have to be considered essential in the all-subsets regression formulas for a given species. The PCA was resorted to in certain situations, for example, when an AMR determinant was not included in any species' formulas, or in some cases for turkey, where all formulas resulted in very low adjr2 values. In cases where it was not possible to attribute one specific reservoir to an AMR determinant based on the all-subsets regression and the PCA, that AMR determinant remained linked to the reservoirs in which it was most abundant.

The results of this study were ultimately compared to those of Duarte et al. (2021), in order to compare the AMR determinants classified as resistome signatures in both studies, and also to verify differences in the methods used for the analyses.

3.4.1. Summary of previous study

The main goal of Duarte et al. (2021) was to predict the sources of resistomes of humans with exposure to livestock and also to identify reservoir-specific resistome signatures. The raw data used by Duarte et al. (2021) was the same used to carry out the present work.

Two analyses were performed. One was to fit a total of 5 random forest models, assuming both a hierarchical and a flat structure of the data, and using different data subsets for model training (see further details of methods in (Duarte et al. 2021). The other was a similarity percentage analysis (Simper), represented in Figure 5, which makes pairwise comparisons of the AMR determinants and calculates the contribution of each determinant to the dissimilarity between every 2 reservoirs. Based on this analysis, Duarte et al. (2021) considered 16 AMR determinants as signatures. The criterion was that if a determinant was in the top ten of determinants that contributed more than 50% for the dissimilarity between two reservoirs and it was always more abundant in one of the reservoirs, it was considered a signature for the reservoir it was most abundant in.

Some of the top contributors were identified as resistome signatures of exclusively one reservoir. These included the determinants *aph(3'')-II*, *aph(3'')-IIIa* and *sul2* as markers of veal calves resistomes, *tet(A)*, *tet(L)* and *tet(S/M)* as markers of turkey resistomes, *ant(6)-Ia* as marker of pig resistomes, *lnu(A)* as marker of broiler resistomes, and *cfxA6* as marker of

human resistomes. Further, determinants *erm(F)*, *mef(A)*, *tet(40)* and *tet(Q)* were identified as resistome signatures primarily in veal calves and secondarily in pigs, while determinants *bla$_{TEM}$-126* and *erm(B)* as resistome signatures primarily in broilers or turkeys. Similarly, determinant *cfxA2* was identified as a marker primarily of veal calves and secondarily of human resistomes.

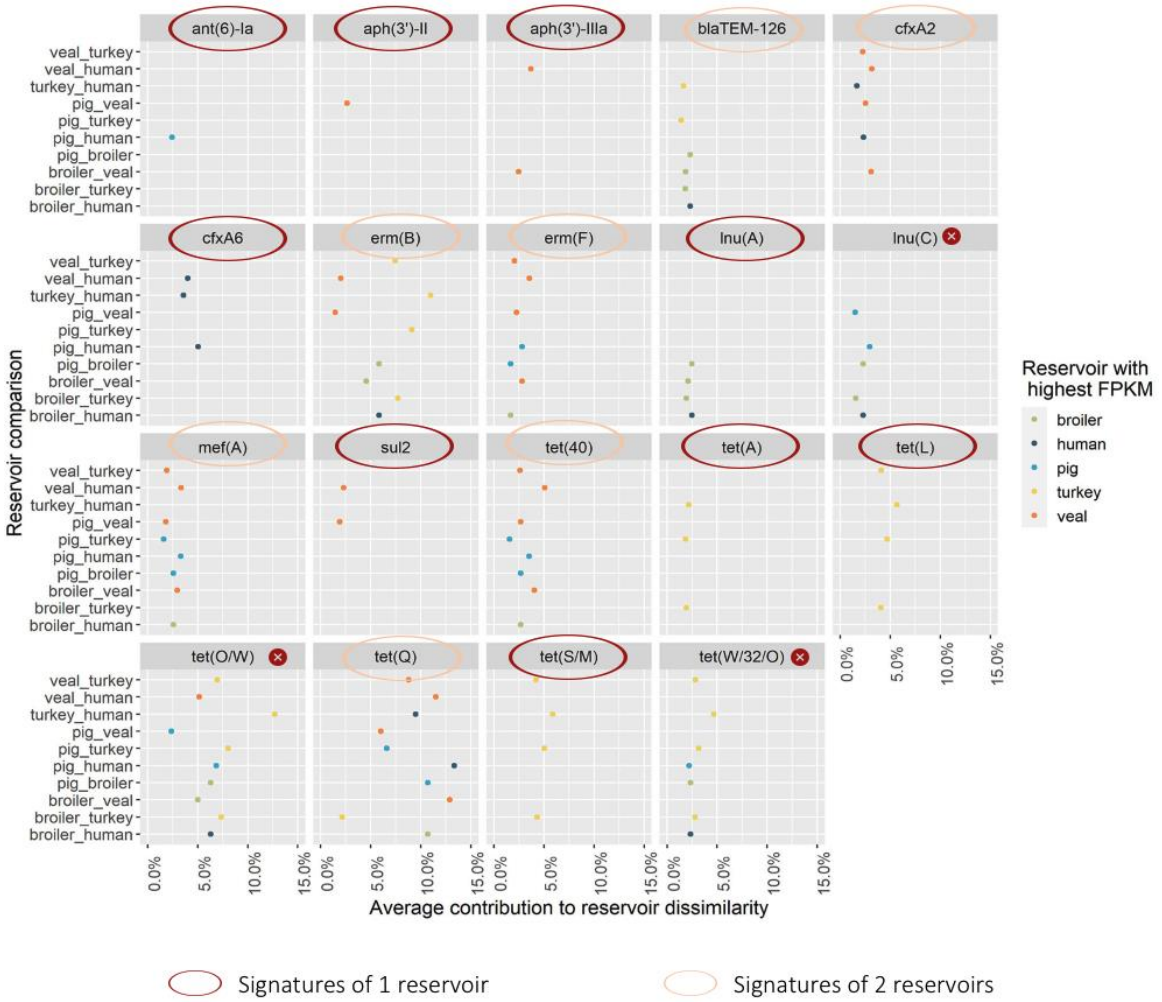

Figure 5 – Simper analysis of the previous study.

AMR determinants with the top contribution to pairwise dissimilarity between resistomes of five reservoirs. The color of the points represents the reservoir for which the AMR determinant had a higher relative abundance (FPKM) in a pairwise comparison. The horizontal axis represents the average proportional contribution of an AMR determinant to the overall average dissimilarity between two reservoirs. The vertical axis shows the two reservoirs compared. AMR determinants in red circles are signatures for a single species (for example *ant(6)-Ia* is a resistome signature for pigs), while the AMR determinants marked in orange are those considered signatures primarily in one reservoir and secondarily in another (for example *erm(F)* was considered a signature primarily of veal calves and secondly of pigs). AMR determinants marked with a red circle with cross were not considered resistome signatures of any reservoir. [The use of these results within the scope of this project was authorized by the authors of Duarte et al. (2021)]

4. CHAPTER IV – RESULTS

4.1. Random forest model

4.1.1. Measures of performance of the model

After running the model with the training set, a summary of the final model's features and performance metrics was obtained, presented in Figure 6. The number of random variables used at each split (mtry) was 30, and the final model's accuracy was very high, with a value of 0.98. The confusion matrix of the OOB predictions shows that the model incorrectly classified 9 of the 384 samples, 7 turkey samples classified as broiler and 2 veal samples classified as pig, resulting in an OOB error of 2.34%.

```
## Call:
##  randomForest(x = x, y = y, ntree = 200, mtry = min(param$mtry, ncol(x)))
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 30
##
##         OOB estimate of  error rate: 2.34%
## Confusion matrix:
##         pig broiler turkey veal class.error
## pig     145       0      0    0  0.00000000
## broiler   0     142      0    0  0.00000000
## turkey    0       7     41    0  0.14583333
## veal      2       0      0   47  0.04081633
##    mtry      Accuracy
##      30     0.9812800
```

Figure 6 – Final model summary performance after training.

Similarly to training, a summary of the model's performance was obtained, after making predictions with the testing set, which is presented in Figure 7. The model only predicted incorrectly 2 of the 95 samples of the testing set, specifically 2 turkey samples which were classified as broiler. The RF model reached an accuracy rounded value of 0.98, very similar to the training performance. The MCC, which is considered more realistic, reached a rounded value of 0.97.

Overall, the performance results were very positive and demonstrated that the model was capable of correctly attributing different classes to observations based on the abundance levels of various AMR determinants.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction pig broiler turkey veal
##    pig      36       0      0    0
##    broiler   0      35      2    0
##    turkey    0       0     10    0
##    veal      0       0      0   12
##
## Overall Statistics
##
##                Accuracy : 0.9789
##     P-Value [Acc > NIR] : < 2.2e-16
##
##          MCC
## 1 0.9698665
```

Figure 7 – Model performance on testing set.

4.1.2. Evaluation of important variables for the model

The built-in variable importance measure of RF resulted in a ranking of the top 15 variables, which can be seen in Figure 8. Although a descriptive analysis was performed on all 15 important variables presented, only 9 variables demonstrated to have high importance (above 50). These included, in abbreviated names, the beta-lactam resistance determinants *cfxA6*, *blaACI* and *cfxA2*, the tetracycline resistance determinants *tet(W)*, *tet(Q)* and *tet(40)*, the macrolide resistance determinants *mef(A)_3* and *mdf(A)*, and the sulphonamide resistance determinant *sul2*. Among those, *cfxA6* reached an importance level of 100 and *blaACI* an importance level of 96, thus contributing together to most prediction accuracy of the model.

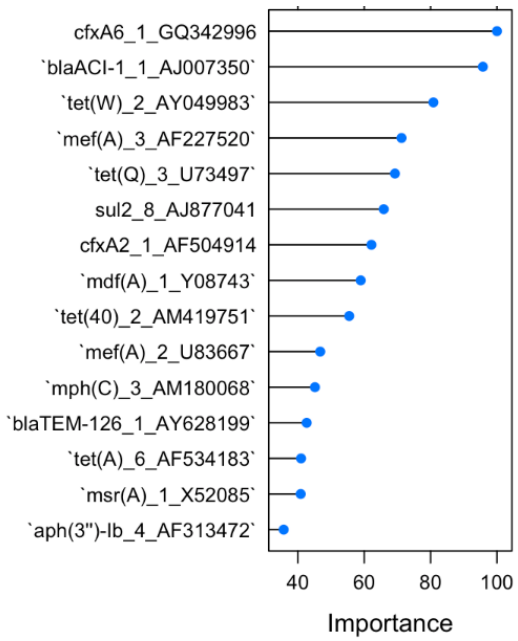

Figure 8 – Ranking of 15 most important variables.

4.1.3. Testing model's reliability

4.1.3.1. Testing with shuffled data and zero important variables

The results of testing the model with shuffled data and zero important variables are presented in Figure 9. The main aim was to compare the performances of the test models (Figure 9) with the performance of the original model (Figure 7), with particular attention to the MCC.

When applied with **shuffled data**, the model incorrectly classified 63 of the 95 observations in the testing set, resulting in an accuracy rounded value of 0.34. The MCC value was especially low, reaching -0.06.

When applied with only **15 zero-importance variables**, the model incorrectly classified 56 of the 95 observations, which resulted in an accuracy value of 0.41. The MCC was 0.11, also extremely low. These results indicate that the variable importance scores in the original model were not being attributed randomly.

Performance with shuffled data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction pig  broiler turkey veal
##    pig      17      20      8    5
##    broiler  19      15      4    7
##    turkey    0       0      0    0
##    veal      0       0      0    0
##
## Overall Statistics
##
##                Accuracy : 0.3368
##     P-Value [Acc > NIR] : 0.8292
##
##           MCC
## 1 -0.06334073
```

Performance with zero important variables

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction pig  broiler turkey veal
##    pig      36      32      9   12
##    broiler   0       3      3    0
##    turkey    0       0      0    0
##    veal      0       0      0    0
##
## Overall Statistics
##
##                Accuracy : 0.4105
##     P-Value [Acc > NIR] : 0.2965
##
##         MCC
## 1 0.1129411
```

Figure 9 – Model's performance with shuffled data and zero important variables.

4.1.3.2. Tests series

The **first test series** demonstrated that adding the 15 most important variables one at a time caused a progressive increase in the MCC value, as can be seen in Figure 10.

Firstly, a model was executed using only the most important AMR determinant (*cfxA6*), which resulted in an MCC of 0.56, a reasonable value considering that the model used the abundance values of only one determinant. As expected, as more important variables were included as predictors, the model had access to more information and was increasingly capable of classifying observations based on those patterns, increasing the value of MCC.

MCC values increased suddenly when the *mef(A)_3* determinant was added and then when the *mdf(A)* determinant was added, after which they reached a plateau. The performance of the model with the 15 most important variables resulted in an MCC of 0.95, almost as high as the MCC of the original model, which used all 389 variables.

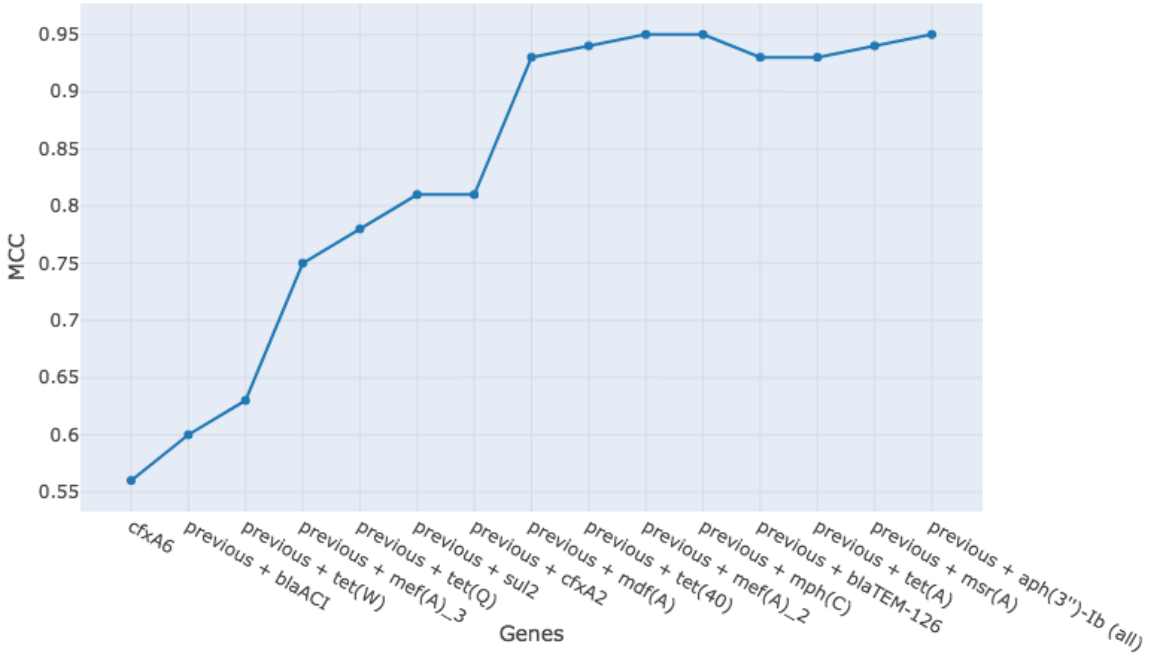

Figure 10 – Test series using an increasing number of the 15 most important variables.
Variables were added in descending order of importance, from most (cfxA6) to least important (aph(3'')-Ib).

The **second test series** (Figure 11) used random AMR determinants without including the 15 most important ones. It was able to provide information about the nuances of the model's performance and the important variables' ranking, confirming previous results.

Figure 11 – Test series using an increasing number of random variables.

The performance of the model using only one random variable resulted in an MCC of 0.14, which is an extremely low value, almost as low as the test models' performance with shuffled data and with zero important variables (Figure 9). In fact, when the model used 4 variables, it reached -0.1, even lower than the MCC with shuffled data.

Overall, the MCC values were very irregular when the model used only a few random variables, increasing more regularly from the moment that 8 to 10 variables are used. The performance of the model with 15 random variables reached a maximum of 0.6, a value that is not representative of an accurate model, comparing to the first test series.

4.2. Descriptive analysis

4.2.1. Detailing characteristics of important AMR determinants

The previous results showed the RF model's accuracy to be reliable, confirming the 15 AMR determinants as essential.

Detailed information about each important determinant is presented in Table 3, such as the importance value in the RF model, the abbreviated name of the cluster's representative gene, the cluster number, and the class of antimicrobials to which it confers resistance.

Table 3 – Details of 15 important AMR determinants.

Importance	Cluster name	Abbreviation	Cluster number	Antimicrobial class
100.00	cfxA6_1_GQ342996	cfxA6	177	Beta-lactam
95.77	blaACI-1_1_AJ007350	blaACI	308	Beta-lactam
80.83	tet(W)_2_AY049983	tet(W)	34	Tetracycline
71.24	mef(A)_3_AF227520	mef(A)_3	114	Macrolide
69.26	tet(Q)_3_U73497	tet(Q)	21	Tetracycline
65.86	sul2_8_AJ877041	sul2	316	Sulphonamide
62.16	cfxA2_1_AF504914	cfxA2	189	Beta-lactam
58.94	mdf(A)_1_Y08743	mdf(A)	108	Macrolide
55.46	tet(40)_2_AM419751	tet(40)	113	Tetracycline
46.73	mef(A)_2_U83667	mef(A)_2	116	Macrolide
45.14	mph(C)_3_AM180068	mph(C)	229	Macrolide
42.65	bla_{TEM}-126_1_AY628199	bla_{TEM}-126	293	Beta-lactam
40.95	tet(A)_6_AF534183	tet(A)	96	Tetracycline
40.85	msr(A)_1_X52085	msr(A)	80	Macrolide
35.71	aph(3")-Ib_4_AF313472	aph(3")-Ib	400	Aminoglycoside

Figure 12 illustrates the number of important AMR determinants belonging to each antimicrobial class. The majority of determinants confer resistance to antimicrobials belonging to the beta-lactam, tetracycline and macrolide classes. Only determinant *sul2* and determinant *aph(3'')-Ib* confer resistance to sulphonamides and aminoglycosides, respectively.

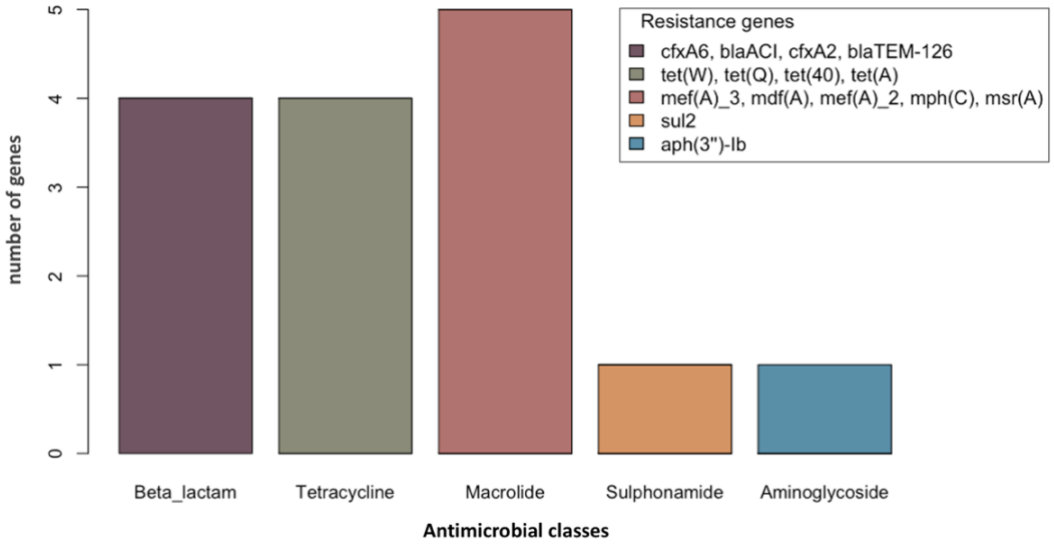

Figure 12 – Antimicrobial classes of 15 important AMR determinants.

The distribution of relative abundance (in FPKM) of each AMR determinant across different species is presented in Figure 13, which provides further insights into the model's results.

Firstly, the independent length of each y-axis must be taken into account, as the abundance values of the 15 important AMR determinants reach very different values. For example, the *tet(Q)* determinant reaches abundance values above 400, while the *tet(W)* determinant, despite being more important and having similar abundance patterns, only reaches FPKM values of around 10.

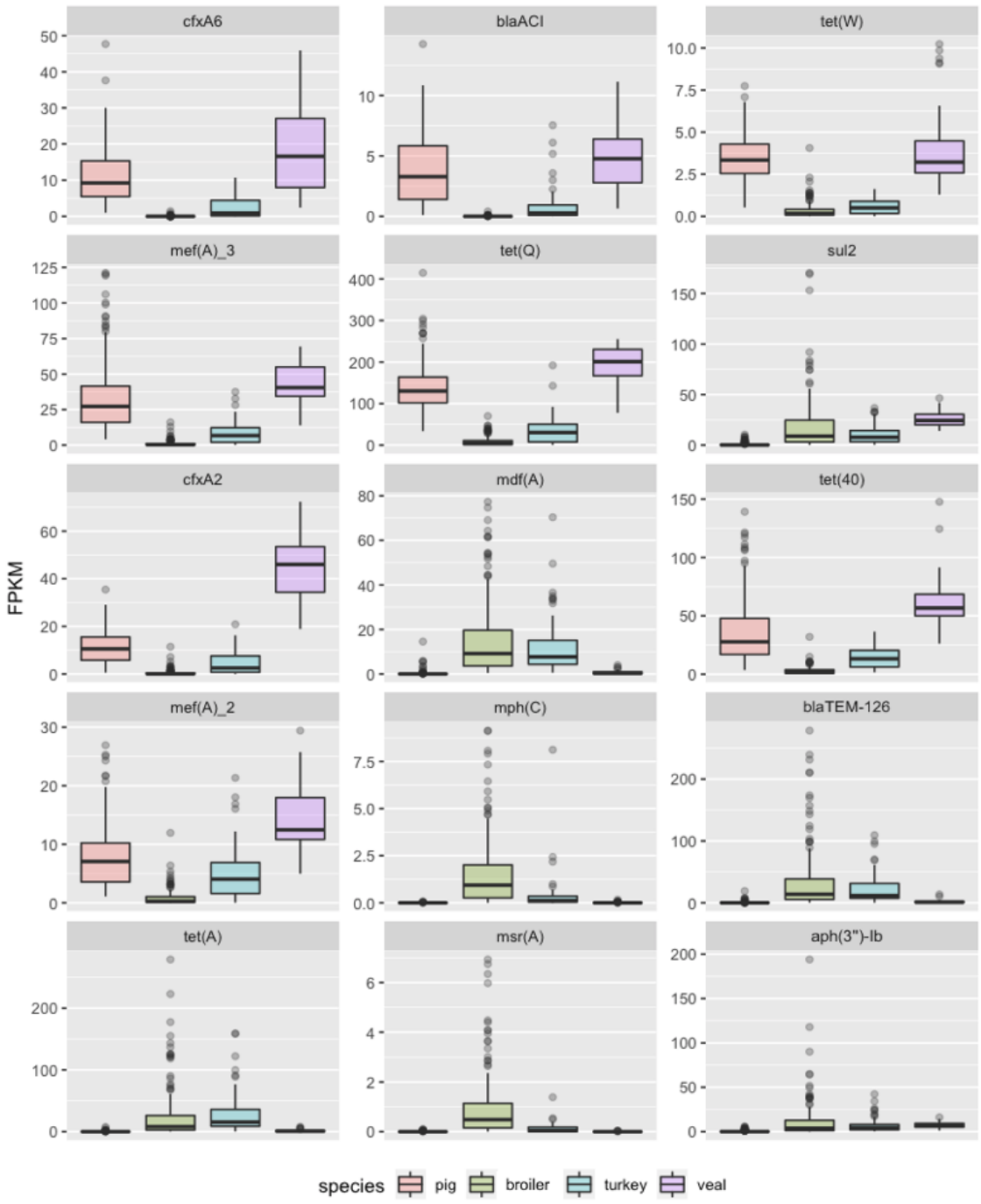

Figure 13 – Distribution of FPKM values of 15 AMR determinants among each reservoir.

Overall, it is possible to observe some patterns of abundance. Most AMR determinants tend to be most abundant in either pig and veal or broiler and turkey. Specifically, the *cfxA6*, *blaACI*, *tet(W)*, *mef(A)_3*, *tet(Q)*, *cfxA2*, *tet(40)* and *mef(A)_2* determinants are more abundant in pig and veal, with low values in broiler and turkey, while the *mdf(A)*, *mph(C)*, bla_{TEM}*-126*, *tet(A)* and *msr(A)* determinants are more abundant in broiler and turkey (mostly in broiler), with low or null values in pig and veal.

This is, in general, in agreement with the model's classification, which showed greater difficulty in distinguishing pig from veal, and broiler from turkey.

4.2.2. Principal Component Analysis

Each PCA resulted in two plots. The scores plot demonstrates how the samples cluster in the newly created dimensions, and the loading plot shows which variables are most responsible for the data variance in these same dimensions. Since the first two dimensions explain most of the data variance, only the plots for these dimensions are presented here, in Figures 14 and 15. Also, to better distinguish the variables' identity in the loading plots, the clusters' numbers, shown in Table 2, were used, instead of abbreviation name.

The **first PCA** (Figure 14) was made using the 15 AMR determinants important for the model. According to the scores plot, it is possible to separate the samples into four different groups, corresponding to the four different species. The first dimension (Dim1 or PC1) explains 72.5% of all the data variance, and mostly splits the data into two major groups along this axis, pig-veal and broiler-turkey, which is an indication of how different these two groups are. The pig and veal clusters range greatly along Dim1 but do not range much along the second dimension (Dim2 or PC2), with values close to zero. Broiler and turkey samples clustered with great range along the Dim2 axis, with coordinates that largely overlap, and to a lesser extent along the negative coordinates of Dim1 axis. Although the broiler cluster ranges greatly along Dim2, it is important to note that this dimension only captures 17.3% of all data variance. Since Dim1 explains most of data variance, the way pig and veal clusters spread along this axis indicate that the abundance values of pig and veal samples contain greater variance than broiler and turkey samples.

In the loading plot it is possible to verify that the determinants 21 (*tet(Q)*), 113 (*tet(40)*), 114 (*mef(A)_3*), 189 (*cfxA2*) and 177 (*cfxA6*) influence the variance of the pig and veal samples. The farther from the origin, the more influence on the samples, indicating that determinant 21 (*tet(Q)*) was the most responsible for the pig and veal data variance. On the other hand, determinants 293 (bla_{TEM}-*126*), 96 (*tet(A)*), 316 (*sul2*), 108 (*mdf(A)*) and 400 (*aph(3")-Ib*) influenced the variance in the broiler and turkey samples, with determinants 293 (bla_{TEM}-*126*) and 96 (*tet(A)*) being the main influencers of broiler and turkey data variance and the main responsible determinants for the outliers of broiler in the scores plot, which matches with the outliers observed in the abundance distribution of these AMR determinants in Figure 13.

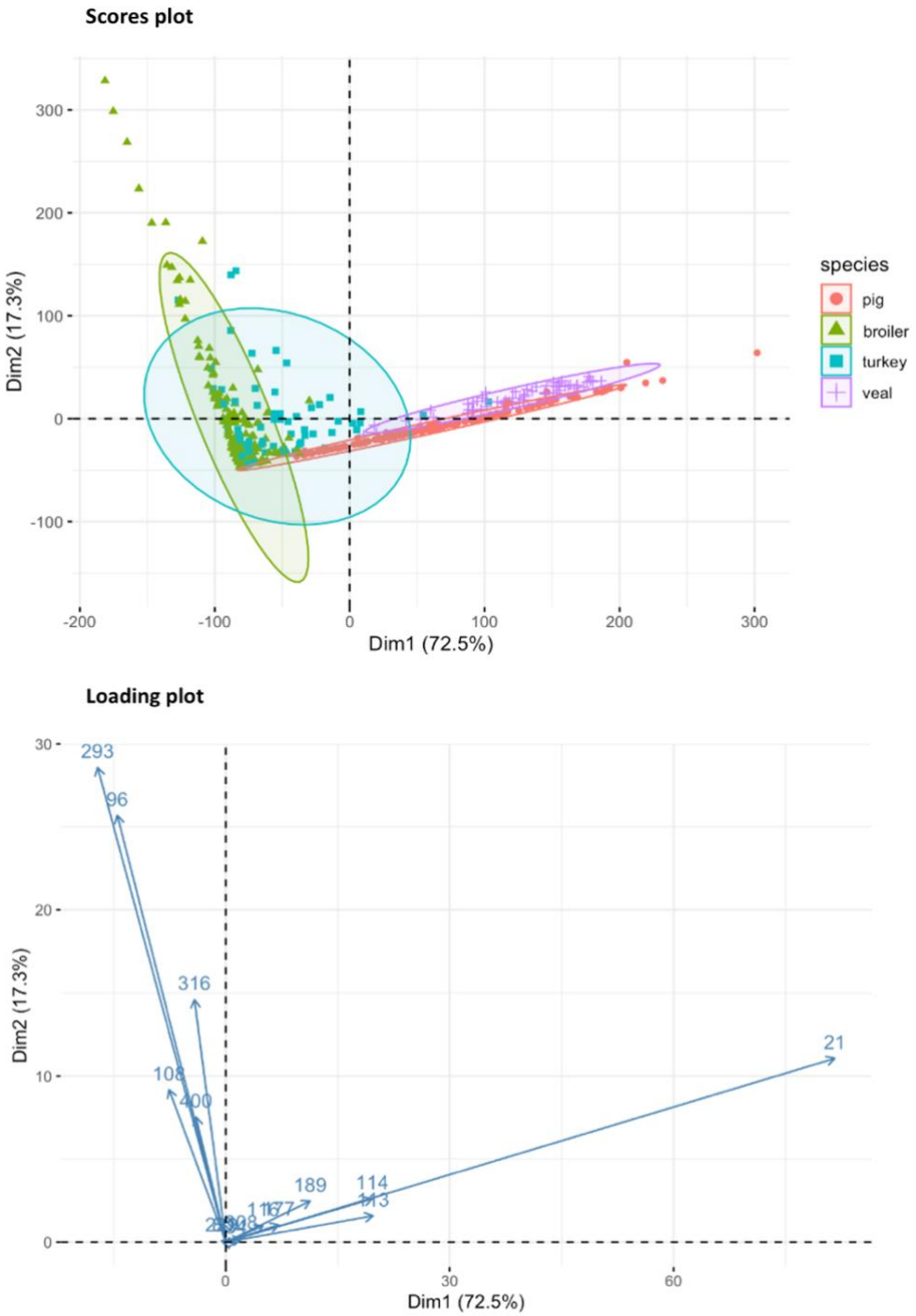

Figure 14 – PCA with 15 important AMR determinants.

The **second PCA** (Figure 15) was performed using all AMR determinants as variables. This time, the clusters appear reversed in the scores plot, with the broiler and turkey clusters more evident along Dim1, and the pig and veal clusters along Dim2. The shape of these clusters remains similar, only with an inverted location. However, this time Dim1 only explains 47.1% of the data variance, which is not surprising given that it has a larger number of variables to explore.

The loading plot shows that determinant 21 (*tet(Q)*) remains the main influencer in pig and veal samples. On the other hand, this analysis demonstrated that there are other AMR determinants influencing broiler and turkey variance, which were not considered important for the model, such as determinants 29 (*tet(O/W)*), 454 (*erm(B)*), 32 (*tet(W/32/O)*), 25 (*tet(S/M)*) and 90 (*tet(L)*). Similarly, determinant 408 (*erm(F)*) was demonstrated to influence pig and veal variance, and determinant 611 (*lnu(A)*) was shown to be responsible for the broiler outliers, along with determinants 96 (*tet(A)*) and 293 (bla_{TEM}-*126*) previously described.

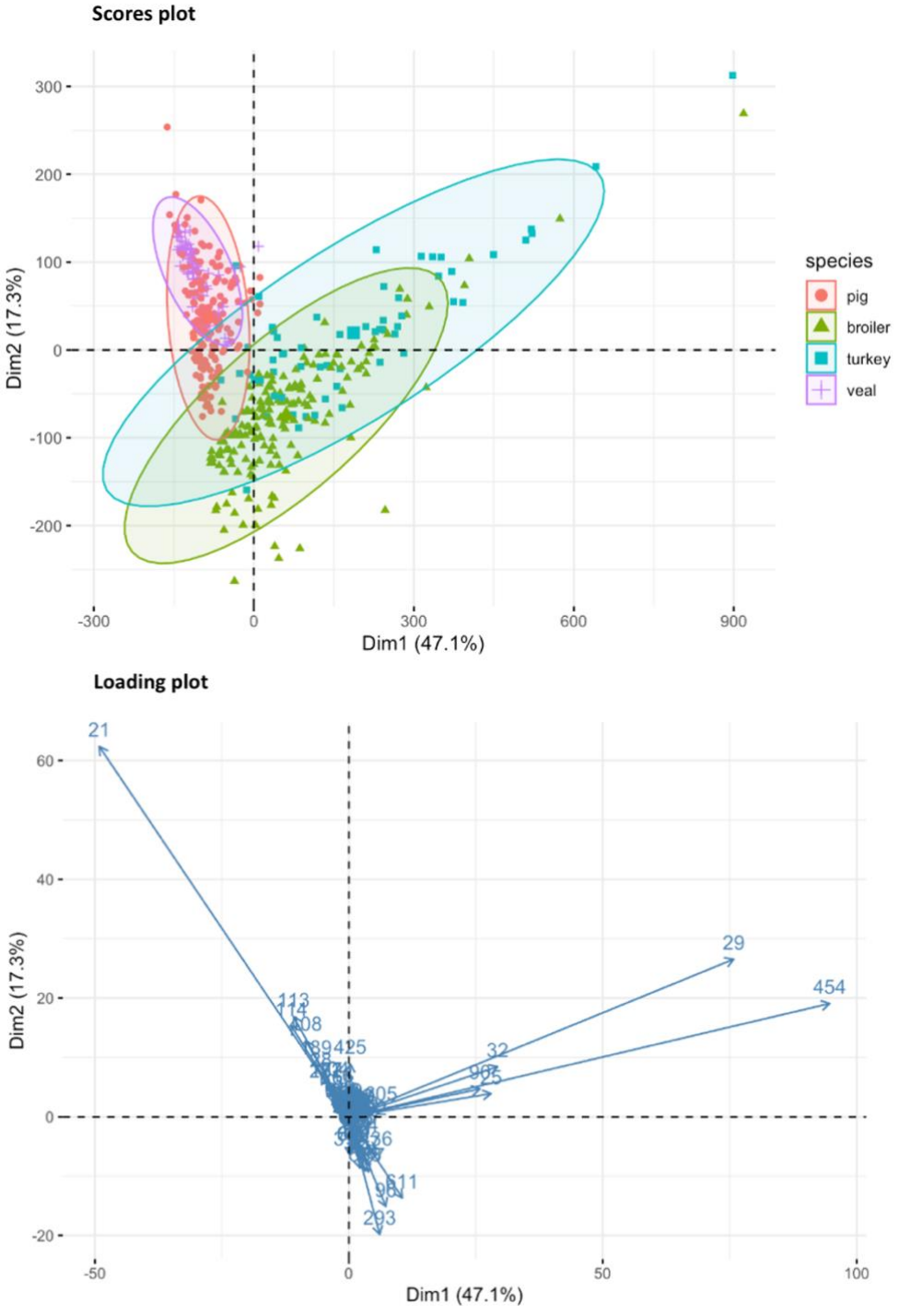

Figure 15 – PCA with all data.

4.2.3. All-subsets regression

An all-subsets regression was performed for each reservoir. The two best formulas (models) were calculated for each number of AMR determinants used in that formula (subsets). Each analysis resulted in a graph for the respective species, presented below.

The first regression was performed for **pig** (left-side graph of Figure 16). Looking at the bottom rows of this graph, a model with only the intercept and determinant *tet(Q)* had an adjusted R-square (adjr2) of 0.26 and a model with only the intercept and determinant *tet(W)* had an adjr2 of 0.36. According to this regression, these were the 2 best formulas to classify a sample as pig using only one determinant. Jumping to the top rows, the formulas with the highest adjr2 values are displayed, suggesting that those are the best formulas. For pig, the best formulas, with adjr2 of 0.77, included the determinants *cfxA6*, *tet(W)*, *mef(A)_3*, *tet(Q)*, *sul2*, *cfxA2*, and *mef(A)_2*. The determinants *blaACI* and *mph(C)* were not considered since the third best formula, without these determinants, also resulted in an adjr2 of 0.77. Considering that determinants *cfxA6*, *tet(W)*, *tet(Q)* and *cfxA2* are most frequently used, even in formulas with a lower number of variables, it is possible to infer that these AMR determinants are responsible for most of relative abundance variance in pig samples.

The right-side graph of Figure 16 shows the best formulas for **veal**, on the top rows, with adjr2 of 0.81. The determinants *blaACI*, *tet(Q)*, *sul2*, *cfxA2* and *tet(40)* appeared to be the most frequently incorporated into formulas to predict veal, indicating to be essential AMR determinants in veal resistomes. Additionally, when compared to pig formulas, veal formulas using a small number of variables (bottom rows) reached very high adjr2 values. For example, the formula with only determinant *mef(A)_2* to predict veal has an adjr2 of 0.28, but the formula using only determinant *cfxA2* reaches an adjr2 of 0.72, indicating that this particular AMR determinant can explain a large part of veal's relative abundance variance.

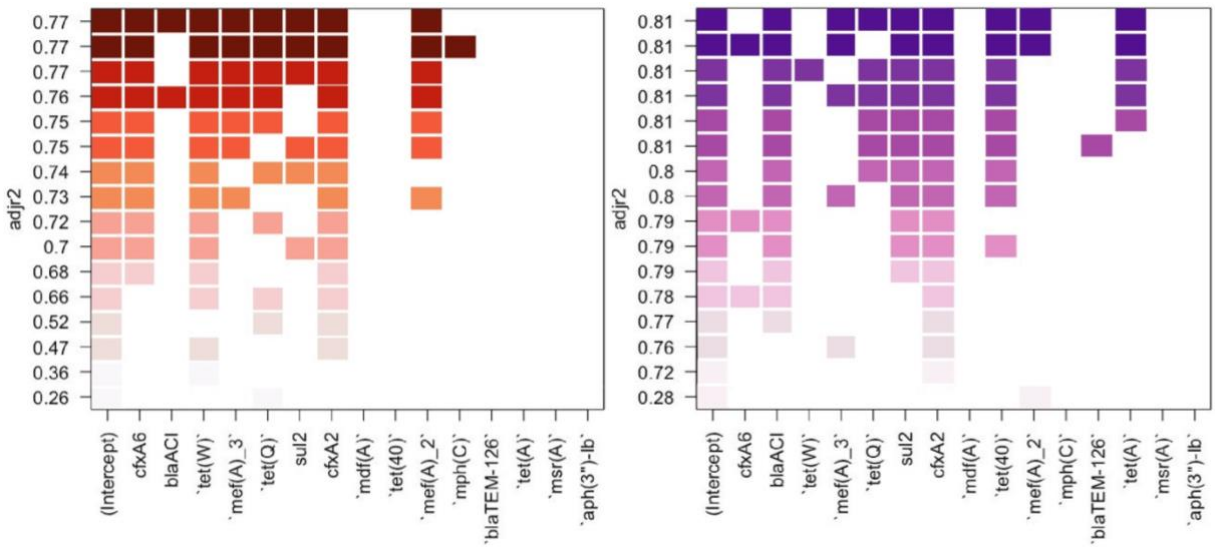

Figure 16 – All-subsets regression for pig (left) and veal (right).

For **broiler** (left-side graph of Figure 17), the best formulas reached 0.64 of adjr2, which is not as high as pig and veal formulas. The most important AMR determinants to predict this reservoir were *tet(W)*, *tet(Q)* and *msr(A)*, since they are the most frequently used in formulas, even in the smaller subsets of variables.

The right-side graph of Figure 17 shows the best formulas for **turkey**. In this case, the most important AMR determinants to predict turkey appear to be *tet(W)*, *mef(A)_3*, *mef(A)_2* and *msr(A)*. However, the maximum adjr2 value that turkey's formulas reach is 0.28. Such low values of adjr2 indicate that these determinants do not explain the dependent variable (turkey species) and are not very reliable to predict this reservoir. In fact, formulas that use smaller subsets of variables reach almost zero values of adjr2, such as when only using the *mef(A)_3* and *mef(A)_2* determinants, or the *tet(W)* and *msr(A)* determinants. Thus, based on the all-subsets regression alone, none of these AMR determinants can be associated with certainty to the reservoir turkey.

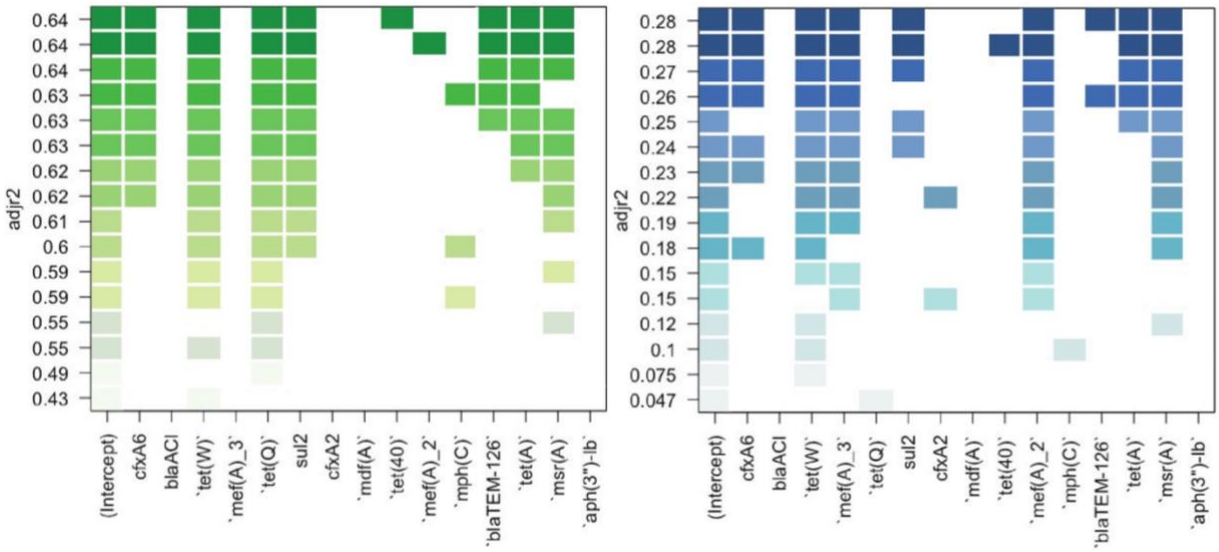

Figure 17 – All-subsets regression for broiler (left) and turkey (right).

4.3. Comparison with Simper analysis

According to the pairwise dissimilarity analysis by Duarte et al. (2021), the highest resistome dissimilarity was between broiler-veal calves and broiler-pig, while pig and veal calves were the two reservoirs with the most similar resistome compositions. The two poultry species' resistomes were less similar than the resistomes of pigs and veal calves.

A comparison of this study's results with the Simper analysis is shown in Table 4. The important AMR determinants for RF model correspond to the Simper signatures in about 50%. Of the 389 AMR determinants, 8 were considered reservoir-specific in both studies. These were determinants *sul2*, *tet(A)*, *cfxA6*, *mef(A)*, *tet(40)*, *tet(Q)*, *cfxA2*, and *bla$_{TEM}$-126*. Most

reservoirs of which these AMR determinants are signatures of, are in agreement in both studies. These results are further discussed in Chapter V.

Table 4 – Comparison between RF reservoir-specific determinants and Simper signatures.

Class	AMR determinants	Analysis	Reservoir RF	Reservoir Simper
Tetracycline	*tet(40)*	**RF + Simper**	veal	veal / pig
	tet(Q)	**RF + Simper**	veal / pig	veal / pig
	tet(A)	**RF + Simper**	broiler / turkey	turkey
	tet(W)	RF	broiler / turkey	-
	tet(L)	Simper	-	turkey
	tet(S/M)	Simper	-	turkey
Macrolide	*mef(A)_2 / mef(A)_3*	**RF + Simper**	veal / pig	veal / pig
	mph(C)	RF	broiler / turkey	-
	mdf(A)	RF	broiler / turkey	-
	msr(A)	RF	broiler / turkey	-
	lnu(A)	Simper	-	broiler
	erm(B)	Simper	-	broiler / turkey
	erm(F)	Simper	-	veal / pig
Beta-lactam	*cfxA6*	**RF + Simper**	pig	human
	cfxA2	**RF + Simper**	veal / pig	veal / human
	bla$_{TEM}$-126	**RF + Simper**	broiler	broiler / turkey
	blaACI	RF	veal	-
Aminoglycoside	*aph(3'')-Ib*	RF	broiler / turkey	-
	ant(6)-Ia	Simper	-	pig
	aph(3')-II	Simper	-	veal
	aph(3')-IIIa	Simper	-	veal
Sulpho-namide	*sul2*	**RF + Simper**	veal	veal

5. CHAPTER V: DISCUSSION

Health research has increasingly benefited from supervised classification algorithms over the years, particularly AMR research due to the facilitated access to AMR datasets (Elyan et al. 2022). Random Forests specifically have been demonstrated to be one of the most useful methods for genomic data analysis, being applied in an increasing number of studies (Gupta et al. 2019; Duarte et al. 2021). In this study, a Random Forest model was applied to a genomic AMR dataset with the purpose of attributing the source of AMR determinants to different animal reservoirs.

5.1. Random Forest limitations

RF's algorithm is known for its high accuracy, interpretability, and ability to learn from extensive data. Despite its popularity, RF has some limitations which can lead, in specific scenarios, to suboptimal performances. Here, those limitations are discussed, and potential solutions are explored to address them.

A random forest model can be underfitted or overfitted. The main limitation to the use of RF, and the one that requires greater caution, is its tendency to overfit. Overfitting is a critical problem when training machine learning models, and RF is not an exception (Breiman 2001). It occurs when a model learns the training data too well, in a way that it memorizes its noise and idiosyncrasies, instead of learning the underlying pattern. When a model overfits with a narrow sample that is not representative of the entire population, it will not be able to make accurate predictions with new observations, meaning it will not be generalizable to new datasets (Chen and Ishwaran 2012; Luan et al. 2020).

There are many causes for overfitting to occur. Sample size is one important cause of overfitting. The data used to apply the RF is restricted to 479 observations, which is relatively low. A low number of samples limits the model's learning content which may cause overfitting.

Additionally, the data used in the present study is significantly unbalanced. Of the 479 observations, almost 75% corresponded to pig and broiler samples (181 pig samples, 177 broiler samples, 60 turkey samples, and 61 veal samples). Unbalanced data is one of the biggest challenges when it comes to RF models, and can also cause overfitting. It has been shown that the dominance of majority class observations causes biased performance in RF models, due to its goal of maximizing the overall prediction accuracy (Vuttipittayamongkol et al. 2021). A solution to this problem could be upsampling the data, to guarantee the proportion of same class samples. However, this is done by replicating the minority samples, and some studies suggest that oversampling very unbalanced data can also lead to overfitting, due to the similarity between same class samples and low diversity of patterns (Elyan et al. 2022). For this reason, the data was kept without upsampling. To make sure the model is reliable and

generalizable, the number of samples collected from each reservoir should be significantly higher and balanced.

Model tuning can also cause overfitting if done incorrectly. When model parameters are too tuned to the training set, the model becomes too complex and specific for those samples (Ghojogh and Crowley 2019 May 28). In a future application of this model to new data, it may be necessary to adjust the tuning parameters (number of trees, depth of trees) for a better generalization. Overall, it is important to be mindful of these factors when training a RF model, and to strike for balance between model complexity and generalization performance.

Another major limitation of RF is its incapability to capture linear relationships. RF models are very effective at capturing the complex, non-linear relationship in the data. However, they are not suitable for capturing linear relationships between input features and target variables because they rely on non-linear boundaries (Louppe 2014). Hence, it was essential to complement this study with more linear methods, such as PCA and All-subsets regression.

5.2. Making sense of resistome signatures

In this study, it was possible to identify important AMR determinants, linking the resistance to certain antimicrobials with specific animal reservoirs. A significant challenge of implementing classification algorithms on metagenomic data is determining the set of features that yields the best discriminatory level for future predictions, especially when features are not present in a single source only. The RF model applied in this study was useful in predicting the sources of a large group of AMR determinants and finding the set of features most responsible for that prediction, as a first step of feature selection. It is important to note that the RF model used was flat, not taking into account the different countries of origin for the samples. The samples were collected from 9 different countries in an unbalanced way, so if there is an interest in assessing country effect in the future, it may be necessary to use specific models for each country, as each country presents different heterogeneity between species.

Based on the model training results (Figure 6), it was possible to observe that the model mostly failed to differentiate pig from veal and broiler from turkey, indicating that pigs have a resistome composition more similar to veal calves, and broilers have a more similar resistome composition to turkeys. The same happened in the predictions' results, where the model only puzzled with two turkey samples, mistaking them for broilers. These results confirm the general intuition that mammalian species should have a more similar microbiome between them than bird species, and vice versa. Additionally, a possible justification for more misclassifications of turkey and veal samples is that there are fewer samples of these two species. As mentioned previously, when the classes of observations are quite unbalanced, the RF algorithm tends to

be biased in its classifications towards the most represented classes, in this case pigs and broilers, by learning mostly the underlying patterns of those samples.

Testing the RF model with different strategies provided evidence about the model's reliability, thus proving model performance and important variables selection to be trustworthy. Testing with shuffled data and zero-importance variables served as control, aiming to test the model's correct functioning. When tested with shuffled data, the MCC was -0.06, indicating that the model did not learn the data's underlying patterns when the classes were randomly shuffled. If the model turned out to be accurate with the shuffled data, it would mean it was not working correctly. When tested with 15 zero-importance variables, the MCC was 0.11. Both results are positive and in line with expectations. Such low MCC values indicate that the predictions made by these 2 models were no better than random predictions. Comparing with the original model (Figure 7), it is possible to state that the original model is reliable, as well as the ranking of important variables.

The first testing series was performed by adding one important variable at a time, in descending order of importance values. In general, this test series is in agreement with the ranking of important variables, since (1) the first models have reasonable MCC despite using few variables, indicating that by themselves they can provide much useful information to the model, (2) the MCC values increase as more important variables are added, (3) the addition of the last variables, which have lower importance values, does not increase the model's MCC values, leading to a plateau, indicating that the model no longer extracts much information after the addition of the least importants AMR determinants (specifically *mdf(A)*).

The second testing series used an increasing number of random variables. Although the MCC values progressively increase as the model manages to extract more information from several variables, they never reach values as high as those obtained with the important variables. In fact, the highest MCC value that a model with random variables could reach, using 15 AMR determinants, was 0.62, which is not representative of an accurate model, since a model using only the 2 most important AMR determinants, *cfxA6* and *blaACI*, reached similar values. It is important to note that although the random variables selected for this testing set could not include the 15 important variables, they could include variables with zero importance. To further dig into the nuances of model performance, another testing set could be executed, in which none of the 15 important variables and none of the zero-importance variables could be selected. Still, these results indicate that the 15 important variables ranked by the original model are indeed essential for classification among the 4 reservoirs.

The descriptive analysis helped to further investigate the 15 important AMR determinants. Most RF's important determinants express resistance to antimicrobials belonging to the macrolide, tetracycline and beta-lactam classes, followed by a minority of determinants expressing resistance to sulphonamide and aminoglycoside. Duarte et al. (2021)

reached similar conclusions in the Simper analysis, where they demonstrated that these 5 classes were the ones that most contributed to the dissimilarity between resistomes of different reservoirs.

The first PCA was helpful to better understand why those AMR determinants were the most important for the model, by informing about which species they were most related to. All samples clustered clearly, according to their belonging reservoir. Pig and veal samples were shown to contain more variance of relative abundance values than the broiler and turkey samples. Within the two poultry species, broiler samples expressed more variance of abundance values than the turkey samples, presenting many outliers, which could also be seen in the abundance distributions plot (Figure 13). Broiler and turkey clusters, as well as pig and veal clusters, have very similar coordinates in the scores plot, which demonstrates the similarity between broiler-turkey resistomes and pig-veal resistomes. This also indicates that broilers' and turkeys' resistomes are dissimilar from pigs' and veal calves' resistomes, which agrees with findings from previous studies. Munk et al. (2018) calculated the dissimilarities between the gene-level resistomes of pig and broiler samples in the same 9 European countries, and pig samples clustered completely separated from broiler samples, demonstrating that pig and poultry resistomes are very distinct.

The second PCA found variables responsible for data variance, which were not considered important by the model. These were determinants *tet(O/W)*, *erm(B)*, *tet(W/32/O)*, *tet(S/M)*, *tet(L)* and *lnu(A)* for broiler and turkey samples, and determinants *erm(F)* for pig and veal samples.

Upon consideration of all analysis conducted in this study, to each RF's AMR determinant, one or two specific reservoirs were assigned, prioritizing the findings of the all-subsets regression, since it finds linear relationships between variables and reservoir, or alternatively the PCA results.

Ultimately, the present study's results were compared to the previous Simper's results (Table 4). In this study, determinants *cfxA2*, *mef(A)_3*, *mef(A)_2*, and *tet(Q)* were considered veal- and pig-specific. Determinant *cfxA6* was considered pig-specific only, and determinants *blaACI*, *tet(40)* and *sul2* were considered veal-specific only, since they were considered essential only in those reservoirs in the all-subsets regression formulas. Most of the pig-specific determinants belong to determinants conferring resistance to tetracyclines and macrolides. Previous studies have shown that there may be a relationship between antimicrobial resistome and antimicrobial use. Van Gompel et al. (2019) studied this relationship in pig farming in the same 9 European countries, and found particularly positive associations between widely used tetracyclines and macrolides and AMR determinants corresponding to those respective antimicrobial classes.

Duarte et al. (2021) considered *cfxA6* and *cfxA2* resistome signatures of humans instead of pigs. They also found an additional pig signature, *ant(6)-Ia*, and an additional pig and veal signature, *erm(F)*. They also considered *sul2* a resistome signature of veal, along with *aph(3')-II* and *aph(3')-IIIa*. This difference between studies is likely due to the addition of human samples in Duarte et al. (2021). Interestingly, the fact that in this current study the determinants *cfxA6* and *cfxA2* were considered pig-specific endorses another finding by Duarte et al. (2021), which is that humans seem to have a resistome composition closer to pig than to other species.

AMR determinants *aph(3'')-Ib*, *tet(A)*, *tet(W)*, *msr(A)*, *mph(C)*, *mdf(A)* were considered signatures of both broiler and turkey in this study, and bla_{TEM}*-126* of only broiler. AMR determinants *aph(3'')-Ib*, *mph(C)* and *mdf(A)* were not included in the best formulas for these species in the all-subsets regression, but they either were responsible for broiler-turkey variance in the PCA or were more abundant in these species according to the abundance distribution. The all-subsets regression considered the *tet(W)* determinant necessary in pig formulas as well, which raised some doubts about which reservoir this determinant should belong to. Ultimately, because it was part of the best formulas of both poultry species, it was considered broiler- and turkey-specific. However, it is important to allude to the subjectivity on which the criteria for this choice may have been based. Nonetheless, the decision to consider determinant *tet(W)* as broiler and turkey-specific was somewhat validated by previous studies. Dec et al. (2017) detected several tetracycline resistance determinants in broilers, of which *tet(W)* was the most frequently occurring tetracycline resistance determinant, observed in 45% of the isolates. A previous study observed positive associations between beta-lactam, tetracycline, macrolide, and aminoglycoside antimicrobial use in broiler farms and AMR gene clusters conferring resistance to the same class (Luiken et al. 2019). The majority of broiler- and turkey-specific AMR determinants found in this study confer resistance to those antimicrobial classes.

Duarte et al. (2021) considered *tet(A)*, *tet(L)* and *tet(S/M)* resistome signatures of turkey*, lnu(A)* a resistome signature of broiler, while *erm(B)* and bla_{TEM}*-126* were considered resistome signatures of broiler and turkey. This was partially in agreement with the all-subsets regression and the PCA. Additionally, previous studies linked some of these determinants to AMU in broiler farming. Luiken et al. (2019) found a significant positive association between flocks treated with beta-lactams and macrolides and an increased abundance of gene clusters bla_{TEM} and *erm(B)*, and Xiong W, Wang Y, et al. (2018) established that a therapeutic dose of chlortetracycline promoted the increase of tetracycline resistance genes' abundance, specifically determinants *tet(W)* and *tet(A)*.

Some of the pig-specific (*cfxA6* and *cfxA2*) and broiler-specific (*tet(A)* and bla_{TEM}*-126*) resistome signatures were identified in a previous study as part of the core resistome of those

species using the same samples (Munk et al. 2018). The same authors also demonstrated that the total AMU (Antimicrobial Use) from the European Medicines Agency's European Surveillance of Veterinary Antimicrobial Consumption was positively associated with AMR in both pigs and poultry. However, the same study showed that resistomes varied between countries with similarly high and diverse AMU and countries with similarly low AMU, demonstrating that even among European countries the livestock resistomes differ in a country-specific way that might be explained by different AMU patterns (Munk et al. 2018). Further investigations are needed on the association between AMU and AMR using metagenomic data, as well as the relation between AMU and AMR considering the differences among countries.

Of the 15 AMR determinants of RF model, only 8 matched the previous Simper analysis signatures. However, it is important to note that, according to the second PCA (Figure 15), many of the determinants considered mostly responsible for the data variance correspond to the remaining determinants considered resistome signatures by Simper. Specifically, the PCA with all data associated determinant *erm(F)* with pig and veal samples, determinants *tet(O/W)*, *erm(B)*, *tet(W/32/O)*, *tet(S/M)* and *tet(L)* with broiler and turkey samples, and determinant *lnu(A)* to broiler outliers. All of these AMR determinants, except for *tet(O/W)* and *tet(W/32/O)*, match previous findings of the Simper analysis. Perhaps this finding is justified by the fact that both PCA and Simper are linear analyses. It would be interesting to further explore the linear relationship between variables and reservoir classes, in order to understand why the above-mentioned AMR determinants were not considered necessary for classification by the RF model.

Finally, it is essential to bear in mind that the conclusion of this study is based on the results of multiple analyses, complementary to the RF model. The interpretation of several analyses in parallel may imply some subjectivity in the selection of the reservoir-specific AMR determinants.

5.3. Future perspectives

When this work started, there were some expectations in mind that, due to lack of time and unavailable data, were not possible to be fulfilled.

One of these goals was to apply the RF model to metagenomic data from samples of humans without direct exposure to the animal reservoirs. This would allow, not only to test the model with a different data set, but also to identify potential sources of antimicrobial resistance in the general human population. This would further generate more data and more insights on how the scientific community, together with the responsible stakeholders, can ameliorate the strategies on AMR prevention.

Another future purpose of this work would be the development of a source code that would contain all the code needed to apply a Random Forest model like the one described here, as well as the complementary analyses required to identify reservoir-specific AMR determinants. This would facilitate the use of RF models with metagenomic data, making the entire source-attribution process using machine learning methods more accessible, even to non-specialized audiences.

I am grateful for having contributed to the validation of previous studies, and I sincerely hope that this work continues and these future expectations can be materialized.

6. MAIN FINDINGS

In this study, we were able to successfully apply a random forest (RF) model to metagenomic data to determine the source of antimicrobial resistance (AMR) and identify reservoir-specific AMR determinants.

Firstly, the RF model was able to accurately predict the source of AMR in the majority of cases. This was demonstrated through the performance of all tests, which proved the RF model's reliability. We found that the RF model had more difficulty differentiating between certain sources. Specifically, the model had difficulty differentiating between pig and veal, as well as between broiler and turkey, indicating that the pig and veal resistome composition is similar, as well as the broiler and turkey resistome composition.

A descriptive analysis was useful for investigating the 15 most important AMR determinants for prediction. Principal component analysis (PCA) and All-subsets regression helped to determine to which reservoirs those determinants were specifically associated.

Finally, it was possible to validate certain determinants as resistome signatures in previous findings for specific sources. Specifically, determinants *tet(40)* and *sul2* are resistome signatures of veal, determinants *tet(Q)*, *mef(A)* and *cfxA2* are resistome signatures of veal and pig, bla_{TEM}*-126* is a resistome signature of broiler, and *tet(A)* is a resistome signature of broiler and turkey.

In a nutshell, this study demonstrates the potential of machine learning models for source-attribution of AMR using metagenomic data. The findings highlight the importance of complementary analysis in investigating the genetic markers of AMR, and the challenges of differentiating between certain sources. Beyond doubt, this study validates several AMR determinants as resistome signatures for specific animal sources, contributing to inform the development of targeted interventions to prevent AMR.

7. REFERENCES

Aarestrup FM, Wegener HC, Collignon P. 2008. Resistance in bacteria of the food chain: Epidemiology and control strategies. Expert Rev Anti Infect Ther. 6(5):733–750. doi:10.1586/14787210.6.5.733.

Aarestrup FM, Woolhouse MEJ. 2020. Using sewage for surveillance of antimicrobial resistance. Science. 367(6478):630–632. doi:10.1126/science.aba3432.

Abebe E, Gugsa G, Ahmed M. 2020. Review on Major Food-Borne Zoonotic Bacterial Pathogens. J Trop Med. 2020. doi:10.1155/2020/4674235.

Abreu VAC de, Perdigão J, Almeida S. 2021. Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview. Front Genet. 11. doi:10.3389/fgene.2020.575592.

Ahmad MW, Reynolds J, Rezgui Y. 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J Clean Prod. 203:810–821. doi:10.1016/j.jclepro.2018.08.207.

Alsan M, Schoemaker L, Eggleston K, Kammili N, Kolli P, Bhattacharya J. 2015. Out-of-pocket health expenditures and antimicrobial resistance in low-income and middle-income countries: An economic analysis. Lancet Infect Dis. 15(10):1203–1210. doi:10.1016/S1473-3099(15)00149-8.

Aminov RI. 2010. A brief history of the antibiotic era: Lessons learned and challenges for the future. Front Microbiol. 1:134. doi:10.3389/fmicb.2010.00134.

Andrade BB, Reis-Filho A, Barros AM, Souza-Neto SM, Nogueira LL, Fukutani KF, Camargo EP, Camargo LM, Barral A, Duarte Â, et al. 2010. Towards a precise test for malaria diagnosis in the Brazilian Amazon: comparison among field microscopy, a rapid diagnostic test, nested PCR, and a computational expert system based on artificial neural networks. http://www.malariajournal.com/content/9/1/117.

Ballabio D, Grisoni F, Todeschini R. 2018. Multivariate comparison of classification performance measures. Chemometrics and Intelligent Laboratory Systems. 174:33–44. doi:10.1016/j.chemolab.2017.12.004.

Basset E, Keith MS, Armelagos GJ, Martin DL, Villanueva A. 1980. Tetracycline-Labeled Human Bone from Ancient Sudanese Nubia (A.D. 350). Science (1979). 209(4464):1532–1534. doi:doi:10.1126/science.7001623.

Behnke S. 2008. Humanoid Robots-From Fiction to Reality? Künstliche Intell. 22(4):5–9.

Bennett JW, Chung K-T. 2001. Alexander Fleming and the Discovery of Penicillin. Adv Appl Microbiol. 49:163–184. doi:doi:10.1016/s0065-2164(01)49013-7.

Berk RA. 2008. Statistics Statistical Learning from a Regression Perspective. New York: Springer.

Bi Q, Goodman KE, Kaminsky J, Lessler J. 2019. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 188(12):2222–2239. doi:10.1093/aje/kwz189.

Biau G, Devroye L, Lugosi G. 2008. Consistency of Random Forests and Other Averaging Classifiers. Journal of Machine Learning Research. 9:2015–2033. doi:10.1145/1390681.1442799.

Boolchandani M, D'Souza AW, Dantas G. 2019. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 20(6):356–370. doi:10.1038/s41576-019-0108-4.

Boole G. 1854. An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities. [accessed 2022 Nov 28]. https://www.gutenberg.org/files/15114/15114-pdf.pdf.

Breiman L. 1996. Bagging Predictors. Mach Learn. 24:123–140. doi:https://doi.org/10.1007/BF00058655.

Breiman L. 2001. Random forests. Mach Learn. 45(1):5–32. doi:10.1023/A:1010933404324.

Cabello FC, Godfrey HP, Buschmann AH, Dölz HJ. 2016. Aquaculture as yet another environmental gateway to the development and globalisation of antimicrobial resistance. Lancet Infect Dis. 16(7):127–133. doi:10.1016/S1473-3099(16)00100-6.

CDC. 2019 Dec. How resistance happens: Select Germs Showing Resistance Over Time. About Antimicrobial Resistance. [accessed 2022 Oct 5]. https://www.cdc.gov/drugresistance/about/how-resistance-happens.html.

Cecchini M, Langer J, Slawomirski L. 2015. Antimicrobial Resistance in G7 Countries and Beyond: Economic Issues, Policies and Options for Action. [accessed 2022 Jun 27]. https://www.oecd.org/els/health-systems/Antimicrobial-Resistance-in-G7-Countries-and-Beyond.pdf.

Chain E, Florey HW. 1944. Penicillin. Endeavour. January:13–14.

Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. Genomics. 99(6):323–329. doi:10.1016/j.ygeno.2012.04.003.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21(1). doi:10.1186/s12864-019-6413-7.

Chokshi A, Sifri Z, Cennimo D, Horng H. 2019. Global contributors to antibiotic resistance. J Glob Infect Dis. 11(1):36–42. doi:10.4103/jgid.jgid_110_18.

Chopra I, Hesse L, O'Neill A. 2002. Discovery and Development of New Anti-Bacterial Drugs. Pharmacochemistry Library. 32:213–225. doi:https://doi.org/10.1016/S0165-7208(02)80022-8.

Clemente L, Lu F, Santillana M. 2019. Improved real-time influenza surveillance: Using internet search data in eight Latin American countries. JMIR Public Health Surveill. 5(2). doi:10.2196/12214.

Collignon P. 2013. The importance of a one health approach to preventing the development and spread of antibiotic resistance. Curr Top Microbiol Immunol. 366:19–36. doi:10.1007/82_2012_224.

Cowan DA, Ramond JB, Makhalanyane TP, de Maayer P. 2015. Metagenomics of extreme environments. Curr Opin Microbiol. 25:97–102. doi:10.1016/j.mib.2015.05.005.

Cox G, Wright GD. 2013. Intrinsic antibiotic resistance: Mechanisms, origins, challenges and solutions. International Journal of Medical Microbiology. 303(6–7):287–292. doi:10.1016/j.ijmm.2013.02.009.

Cunningham P, Cord M, Delany SJ. 2008. Supervised Learning. In: Cord, M., Cunningham, P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Berlim, Heidelberg: Springer. p. 21–49.

Cutler A, Cutler R, Stevens J. 2012. Random Forests. In: Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. New York, NY: Springer. p. 157–175.

Dadgostar P. 2019. Antimicrobial resistance: implications and costs. Infect Drug Resist. 12:3903–3910. doi:10.2147/IDR.S234610.

Dancer S, Shears P, Platt D. 1997. Isolation and characterization of coliforms from glacial ice and water in Canada's High Arctic. J Appl Microbiol. 82:597–609. doi:10.1111/j.1365-2672.1997.tb03590.x.

Dancer SJ. 2014. Controlling hospital-acquired infection: Focus on the role of the environment and new technologies for decontamination. Clin Microbiol Rev. 27(4):665–690. doi:10.1128/CMR.00020-14.

DANMAP. 2021. Use of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from food animals, food and humans in Denmark. [accessed 2022 Feb 24]. https://www.danmap.org/.

Davies J, Davies D. 2010. Origins and Evolution of Antibiotic Resistance. Microbiology and Molecular Biology Reviews. 74(3):417–433. doi:10.1128/MMBR.00016-10.

D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, et al. 2011. Antibiotic resistance is ancient. Nature. 477(7365):457–461. doi:10.1038/nature10388.

D'Costa VM, McGrann KM, Hughes DW, Wright GD. 2006. Sampling the antibiotic resistome. Science (1979). 311(5759):374–377. doi:10.1126/science.1120800.

Dec M, Urban-Chmiel R, Stępień-Pyśniak D, Wernicki A. 2017. Assessment of antibiotic susceptibility in Lactobacillus isolates from chickens. Gut Pathog. 9(1). doi:10.1186/s13099-017-0203-z.

Dick S. 2019. Artificial Intelligence. Harv Data Sci Rev. 1(1). doi:10.1162/99608f92.92fe150c.

Dietterich TG. 2000. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture notes in Computer Science. Vol. 1857. Berlim, Heidelberg: Springer. http://www.cs.orst.edu/˜tgd.

Dorr PM, Tadesse DA, Zewde BM, Fry P, Thakur S, Gebreyes WA. 2009. Longitudinal study of Salmonella dispersion and the role of environmental contamination in commercial swine production systems. Appl Environ Microbiol. 75(6):1478–1486. doi:10.1128/AEM.01632-08.

Duan Y, Edwards JS, Dwivedi YK. 2019. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. Int J Inf Manage. 48:63–71. doi:10.1016/j.ijinfomgt.2019.01.021.

Duarte ASR, Röder T, Van Gompel L, Petersen TN, Hansen RB, Hansen IM, Bossers A, Aarestrup FM, Wagenaar JA, Hald T. 2021. Metagenomics-Based Approach to Source-Attribution of Antimicrobial Resistance Determinants – Identification of Reservoir Resistome Signatures. Front Microbiol. 11(601407). doi:10.3389/fmicb.2020.601407.

Duarte ASR, Stärk KDC, Munk P, Leekitcharoenphon P, Bossers A, Luiken R, Sarrazin S, Lukjancenko O, Pamp SJ, Bortolaia V, et al. 2020. Addressing Learning Needs on the Use of Metagenomics in Antimicrobial Resistance Surveillance. Front Public Health. 8(38). doi:10.3389/fpubh.2020.00038.

Durand G, Raoult D, Dubourg G. 2018. Antibiotic discovery: History, methods and perspectives. J Antimicrob Agents. 53(4):371–382. doi:10.1016/j.ijantimicag.2018.11.010.

Elyan E, Hussain A, Sheikh A, Elmanama AA, Vuttipittayamongkol P, Hijazi K. 2022. Antimicrobial Resistance and Machine Learning: Challenges and Opportunities. IEEE Access. 10:31561–31577. doi:10.1109/ACCESS.2022.3160213.

Fang H, Wang H, Cai L, Yu Y. 2015. Prevalence of antibiotic resistance genes and bacterial pathogens in long-term manured greenhouse soils as revealed by metagenomic survey. Environ Sci Technol. 49(2):1095–1104. doi:10.1021/es504157v.

FAO, OIE, WHO, UNEP. 2022. Quadripartite Memorandum of Understanding (MoU) signed for a new era of One Health Collaboration. [accessed 2022 Sep 9]. https://www.unep.org/resources/publication/quadripartite-memorandum-understanding-mou-signed-new-era-one-health.

Farid DM, Rahman MZ, Rahman CM. 2011. An Ensemble Approach to Classifier Construction based on Bootstrap Aggregation. Int J Comput Appl. 25(5):975–8887.

Fleming A. 1929. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. Br J Exp Pathol. 10(3):226–236.

Fox J. 2008. Applied Regression Analysis and Generalized Linear Models. 2nd ed. SAGE Publications.

François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J. 2018. An introduction to deep reinforcement learning. Foundations and Trends in Machine Learning. 11(3–4):219–354. doi:10.1561/2200000071.

Frege G. 1879. Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. [accessed 2023 Feb 16]. https://www.informationphilosopher.com/solutions/philosophers/frege/Frege_Begriffsschrift.pdf.

Furuta R, Inoue N, Yamasaki T. 2020. PixelRL: Fully Convolutional Network with Reinforcement Learning for Image Processing. IEEE Trans Multimedia. 22(7):1704–1719. doi:10.1109/TMM.2019.2960636.

Garau G, di Guilmi AM, Hall BG. 2005. Structure-based phylogeny of the metallo-β-lactamases. Antimicrob Agents Chemother. 49(7):2778–2784. doi:10.1128/AAC.49.7.2778-2784.2005.

Gasparetto A. 2016. Robots in history: Legends and prototypes from ancient times to the industrial revolution. In: López-Cajún C, Ceccarelli M. (eds) Explorations in the History of Machines and Mechanisms. History of Mechanism and Machine Science. Vol. 32. Springer. p. 39–49.

Gerner-Smidt P, Besser J, Concepción-Acevedo J, Folster JP, Huffman J, Joseph LA, Kucerova Z, Nichols MC, Schwensohn CA, Tolar B. 2019. Whole genome sequencing: Bridging one-health surveillance of foodborne diseases. Front Public Health. 7. doi:doi:10.3389/fpubh.2019.00172.

Ghojogh B, Crowley M. 2019 May 28. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. arXiv preprint arXiv:1905.12787. [accessed 2022 Dec 3]. http://arxiv.org/abs/1905.12787.

Ghosh S, Laksana E, Morency LP, Scherer S. 2016. Representation learning for speech emotion recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Vol. 08-12-September-2016. International Speech and Communication Association. p. 3603–3607.

Gillings MR, Paulsen IT, Tetu SG. 2017. Genomics and the evolution of antibiotic resistance. Ann N Y Acad Sci. 1388(1):92–107. doi:10.1111/nyas.13268.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. 2010. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. BMC Genet. 11:49. doi:https://doi.org/10.1186/1471-2156-11-49.

Van Gompel L, Luiken REC, Sarrazin S, Munk P, Knudsen BE, Hansen RB, Bossers A, Aarestrup FM, Dewulf J, Wagenaar JA, et al. 2019. The antimicrobial resistome in relation to antimicrobial use and biosecurity in pig farming, a metagenome-wide association study in nine European countries. Journal of Antimicrobial Chemotherapy. 74(4):865–876. doi:10.1093/jac/dky518.

Van Gompel L van, Dohmen W, Luiken REC, Bouwknegt M, Heres L, van Heijnsbergen E, Jongerius-Gortemaker BGM, Scherpenisse P, Greve GD, Tersteeg-Zijderveld MHG, et al. 2020. Occupational exposure and carriage of antimicrobial resistance genes (tetW, ermB) in pig slaughterhouse workers. Ann Work Expo Health. 64(2):125–137. doi:10.1093/annweh/wxz098.

Guizzo E. 2019. By Leaps and Bounds: An exclusive look at how Boston Dynamics is redefining robot agility. IEEE Spectr. 56(12):34–39. https://spectrum.ieee.org/bostondynamics1219.

Gupta S, Arango-Argoty G, Zhang L, Pruden A, Vikesland P. 2019. Identification of discriminatory antibiotic resistance genes among environmental resistomes using

extremely randomized tree algorithm. Microbiome. 7(1). doi:10.1186/s40168-019-0735-1.

Haddawy P, Hasan AHMI, Kasantikul R, Lawpoolsri S, Sa-angchai P, Kaewkungwal J, Singhasivanon P. 2018. Spatiotemporal Bayesian networks for malaria prediction. Artif Intell Med. 84:127–138. doi:10.1016/j.artmed.2017.12.002.

Hall BG, Barlow M. 2004. Evolution of the serine β-lactamases: Past, present and future. Drug Resistance Updates. 7(2):111–123. doi:10.1016/j.drup.2004.02.003.

Harlow H. 2018. Ethical Concerns of Artificial Intelligence, Big Data and Data Analytics. In: European Conference on Knowledge Management, Kidmore End: Academic conferences International Limited. p. 316–323.

Hastie T, Tibshirani R, Friedman J. 2001. The Elements of Statistical Learning. 1st ed. Springer Series in Statistics.

Hastie T, Tibshirani R, Friedman J. 2009. Unsupervised Learning. In: The Elements of Statistical Learning. New York, NY: Springer Series in Statistics.

Haulisah NA, Hassan L, Bejo SK, Jajere SM, Ahmad NI. 2021. High Levels of Antibiotic Resistance in Isolates From Diseased Livestock. Front Vet Sci. 8. doi:10.3389/fvets.2021.652351.

Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. 2019. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health. 7. doi:10.3389/fpubh.2019.00242.

Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, Röder T, Nieuwenhuijse D, Pedersen SK, Kjeldgaard J, et al. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun. 10(1). doi:10.1038/s41467-019-08853-3.

Herlau T, Schmidt MN, Morup M. 2021. Introduction to Machine Learning and Data Mining. Technical University of Denmark.

Hilal W, Gadsden SA, Yawney J. 2022. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Syst Appl. 193. doi:10.1016/j.eswa.2021.116429.

Hillock NT, Merlin TL, Turnidge J, Karnon J. 2022. Modelling the Future Clinical and Economic Burden of Antimicrobial Resistance: The Feasibility and Value of Models to Inform Policy. Appl Health Econ Health Policy. 20(4):479–486. doi:10.1007/s40258-022-00728-x.

Hopfield JJ. 1982. Neural networks and physical systems with emergent collective computational abilities. Proc NatL Acad Sci USA. 79(8):2554–2558. doi:https://doi.org/10.1073/pnas.79.8.2554.

Huang K, Tang J, Zhang XX, Xu K, Ren H. 2014. A comprehensive insight into tetracycline resistant bacteria and antibiotic resistance genes in activated sludge using next-

generation sequencing. Int J Mol Sci. 15(6):10083–10100. doi:10.3390/ijms150610083.

Jaeger S, Juarez-Espinosa OH, Candemir S, Poostchi M, Yang F, Kim L, Ding M, Folio LR, Antani S, Gabrielian A, et al. 2018. Detecting drug-resistant tuberculosis in chest radiographs. Int J Comput Assist Radiol Surg. 13(12):1915–1925. doi:10.1007/s11548-018-1857-9.

Jha A, Dave M, Madan S. 2019. Comparison of Binary Class and Multi-Class Classifier Using Different Data Mining Classification Techniques. In: Proceedings of International Conference on Advancements in Computing & Management (ICACM). p. 894–903.

Johnston IG, Hoffmann T, Greenbury SF, Cominetti O, Jallow M, Kwiatkowski D, Barahona M, Jones NS, Casals-Pascual C. 2019. Precision identification of high-risk phenotypes and progression pathways in severe malaria without requiring longitudinal data. NPJ Digit Med. 2(1). doi:10.1038/s41746-019-0140-y.

Kabaria CW, Molteni F, Mandike R, Chacky F, Noor AM, Snow RW, Linard C. 2016. Mapping intra-urban malaria risk using high resolution satellite imagery: A case study of Dar es Salaam. Int J Health Geogr. 15(1). doi:10.1186/s12942-016-0051-y.

Kaiser L, Babaeizadeh M, Milos P, Osinski B, Campbell RH, Czechowski K, Erhan D, Finn C, Kozakowski P, Levine S, et al. 2019. Model-Based Reinforcement Learning for Atari. In: Model Based Reinforcement Learning for Atari. Conference paper at ICLR 2020. [accessed 2023 Jan 12]. http://arxiv.org/abs/1903.00374.

Kaur P, Sharma M. 2019. Diagnosis of Human Psychological Disorders using Supervised Learning and Nature-Inspired Computing Techniques: A Meta-Analysis. J Med Syst. 43(7). doi:10.1007/s10916-019-1341-2.

Kherif F, Latypova A. 2019. Principal component analysis. In: Machine Learning: Methods and Applications to Brain Disorders. Elsevier. p. 209–225.

Kopp J, Wang GY, Horch RE, Pallua N, Ge SD. 2003. Ancient traditional Chinese medicine in burn treatment: A historical review. Burns. 29(5):473–478. doi:10.1016/S0305-4179(03)00053-6.

Kotsiantis SB. 2013. Decision trees: A recent overview. Artif Intell Rev. 39(4):261–283. doi:10.1007/s10462-011-9272-4.

Kozajda A, Jeżak K, Kapsa A. 2019. Airborne Staphylococcus aureus in different environments—a review. Environmental Science and Pollution Research. 26(34):34741–34753. doi:10.1007/s11356-019-06557-1.

Kuhn M. 2008. Building Predictive Models in R Using the caret Package. J Stat Softw. 28(5):1–26. doi:https://doi.org/10.18637/jss.v028.i05.

Kuhn M, Johnson K. 2013. Applied Predictive Modeling. NY: Springer.

Kumari A, Prem Kumar J, Prakash V, Divya K. 2021. Supervised Learning Algorithms: A Comparison. Kristu Jayanti Journal of Computational Sciences. 1(1):1–12. doi:https://doi.org/10.59176/kjcs.v1i1.1259.

Larsson DGJ, Flach CF. 2022. Antibiotic resistance in the environment. Nat Rev Microbiol. 20(5):257–269. doi:10.1038/s41579-021-00649-x.

Lê S, Josse J, Husson F. 2008. FactoMineR: An R Package for Multivariate Analysis. JSS Journal of Statistical Software. 25(1):1–18. doi:10.18637/jss.v025.i01.

Li S, Song W, Fang L, Chen Y, Ghamisi P, Benediktsson JA. 2019. Deep learning for hyperspectral image classification: An overview. IEEE Transactions on Geoscience and Remote Sensing. 57(9):6690–6709. doi:10.1109/TGRS.2019.2907932.

Ljumovic M, Klar M. 2015. Estimating Expected Error Rates of Random Forest Classifiers- A Comparison of Cross-Validation and Bootstrap. In: 4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro. p. 212–215.

Loecher M. 2020. Unbiased variable importance for random forests. Commun Stat Theory Methods. 51(5):1413–1425. doi:10.1080/03610926.2020.1764042.

Louppe G. 2014. Understanding Random Forests: From Theory to Practice [PhD Dissertation]. Faculty of Applied Sciences, University of Liège. http://arxiv.org/abs/1407.7502.

Lu Y. 2019. Artificial intelligence: a survey on evolution, models, applications and future trends. Journal of Management Analytics. 6(1):1–29. doi:10.1080/23270012.2019.1570365.

Luan J, Zhang C, Xu B, Xue Y, Ren Y. 2020. The predictive performances of random forest models with limited sample size and different species traits. Fish Res. 227. doi:10.1016/j.fishres.2020.105534.

Luiken REC, Van Gompel L, Munk P, Sarrazin S, Joosten P, Dorado-García A, Borup Hansen R, Knudsen BE, Bossers A, Wagenaar JA, et al. 2019. Associations between antimicrobial use and the faecal resistome on broiler farms from nine European countries. Journal of Antimicrobial Chemotherapy. 74(9):2596–2604. doi:10.1093/jac/dkz235.

Lumley T. 2003. The leaps package. [accessed 2022 Apr 7]. cran. r-project. org/doc/packages/leaps. pdf.

Mahabub A, Habib A-ZS Bin, Mondal MRH, Bharati S, Podder P. 2020. Effectiveness of Ensemble Machine Learning Algorithms in Weather Forecasting of Bangladesh. In: Abraham A, Sasaki H, Rios R, Gandhi N, Singh U, Kun M, editors. Innovations in Bio-Inspired Computing and Applications. Switzerland: Springer. p. 267–277. http://www.springer.com/series/11156.

Manyi-Loh C, Mamphweli S, Meyer E, Okoh A. 2018. Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential public health implications. Molecules. 23(4). doi:10.3390/molecules23040795.

Marteinson SC, Lawrence MJ, Taranu ZE, Kosziwka K, Taylor JJ, Green A, Winegardner AK, Rytwinski T, Reid JL, Dubetz C, et al. 2022 Jul 30. Increased use of sanitizers and disinfectants during the COVID-19 pandemic: identification of antimicrobial chemicals and considerations for aquatic environmental contamination. Environmental Reviews. doi:10.1139/er-2022-0035.

Martínez-Agüero S, Mora-Jiménez I, Lérida-García J, Álvarez-Rodríguez J, Soguero-Ruiz C. 2019. Machine learning techniques to identify antimicrobial resistance in the intensive care unit. Entropy. 21(6). doi:10.3390/e21060603.

McCarthy J, Minsky M, Rochester N, Shannon C. 2006. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Mag. 27(4):12. doi:https://doi.org/10.1609/aimag.v27i4.1904.

McCorduck P, Cfe C. 2004. Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. 2nd Edition. New York.

Miles J. 2005. R-squared, Adjusted R-squared. In: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons.

Miller A. 2002. Subset Selection in Regression. 2nd Edition. New York: Chapman and Hall/CRC.

Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. 2018. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput Biol. 14(12). doi:10.1371/journal.pcbi.1006258.

Morrison L, Zembower TR. 2020. Antimicrobial Resistance. Gastrointest Endosc Clin N Am. 30(4):619–635. doi:10.1016/j.giec.2020.06.004.

Moyo S, Doan TN, Yun JA, Tshuma N. 2018. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Hum Resour Health. 16(1). doi:10.1186/s12960-018-0329-1.

Munita JM, Arias CA. 2016 May. Mechanisms of Antibiotic Resistance. Microbiol Spectr. doi:10.1128/microbiolspec.vmbf-0016-2015.

Munk P, Andersen VD, de Knegt L, Jensen MS, Knudsen BE, Lukjancenko O, Mordhorst H, Clasen J, Agersø Y, Folkesson A, et al. 2017. A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. Journal of Antimicrobial Chemotherapy. 72(2):385–392. doi:10.1093/jac/dkw415.

Munk P, Knudsen BE, Lukjacenko O, Duarte ASR, Van Gompel L, Luiken REC, Smit LAM, Schmitt H, Garcia AD, Hansen RB, et al. 2018. Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. Nat Microbiol. 3(8):898–908. doi:10.1038/s41564-018-0192-9.

Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, Han C, Bisignano C, Rao P, Wool E, et al. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 399(10325):629–655. doi:10.1016/S0140-6736(21)02724-0.

Nair DVT, Venkitanarayanan K, Johny AK. 2018. Antibiotic-resistant Salmonella in the food supply and the potential role of antibiotic alternatives for control. Foods. 7(10). doi:10.3390/foods7100167.

Nasteski V. 2017. An overview of the supervised machine learning methods. HORIZONSB. 4:51–62. doi:10.20544/horizons.b.04.1.17.p05.

Naylor NR, Atun R, Zhu N, Kulasabanathan K, Silva S, Chatterjee A, Knight GM, Robotham J v. 2018. Estimating the burden of antimicrobial resistance: a systematic literature review. Antimicrob Resist Infect Control. 7:58. doi:10.1186/s13756-018-0336-y.

Nelson ML, Dinardo A, Hochberg J, Armelagos GJ. 2010. Brief communication: Mass spectroscopic characterization of tetracycline in the skeletal remains of an ancient population from Sudanese Nubia 350-550 CE. Am J Phys Anthropol. 143(1):151–154. doi:10.1002/ajpa.21340.

Ng PC, Kirkness EF. 2010. Whole genome sequencing. Methods in Molecular Biology. 628:215–226. doi:10.1007/978-1-60327-367-1_12.

Nguyen M, Wesley Long S, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davisa JJ. 2019. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. J Clin Microbiol. 57(2). doi:10.1128/JCM.01260-18.

O'Neill J. 2016. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Review on Antimicrobial Resistance. UK. [accessed 2022 Nov 22]. https://wellcomecollection.org/works/thvwsuba.

Oonsivilai M, Mo Y, Luangasanatip N, Lubell Y, Miliya T, Tan P, Loeuk L, Turner P, Cooper BS. 2018. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. Wellcome Open Res. 3. doi:10.12688/wellcomeopenres.14847.1.

Oz T, Guvenek A, Yildiz S, Karaboga E, Tamer YT, Mumcuyan N, Ozan VB, Senturk GH, Cokol M, Yeh P, et al. 2014. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. Mol Biol Evol. 31(9):2387–2401. doi:10.1093/molbev/msu191.

Palma E, Tilocca B, Roncada P. 2020. Antimicrobial resistance in veterinary medicine: An overview. Int J Mol Sci. 21(6). doi:10.3390/ijms21061914.

Pećanac M, Zlata J, Komarčević A, Pajić M, Dobanovački D, Mišković S. 2013. Burns treatment in ancient times. Hist Med. 66(5):263–267.

Pérez-Rave JI, Correa-Morales JC, González-Echavarría F. 2019. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research. 36(1):59–96. doi:10.1080/09599916.2019.1587489.

Petersen A, Andersen JS, Kaewmak T, Somsiri T, Dalsgaard A. 2002. Impact of integrated fish farming on antimicrobial resistance in a pond environment. Appl Environ Microbiol. 68(12):6036–6042. doi:10.1128/AEM.68.12.6036-6042.2002.

Petersen TN, Rasmussen S, Hasman H, Carøe C, Bælum J, Charlotte Schultz A, Bergmark L, Svendsen CA, Lund O, Sicheritz-Pontén T, et al. 2015. Meta-genomic analysis of toilet waste from long distance flights; A step towards global surveillance of infectious diseases and antimicrobial resistance. Sci Rep. 5. doi:10.1038/srep11444.

Phakhounthong K, Chaovalit P, Jittamala P, Blacksell SD, Carter MJ, Turner P, Chheng K, Sona S, Kumar V, Day NPJ, et al. 2018. Predicting the severity of dengue fever in

children on admission based on clinical features and laboratory indicators: Application of classification tree analysis. BMC Pediatr. 18(1). doi:10.1186/s12887-018-1078-y.

Pulingam T, Parumasivam T, Gazzali AM, Sulaiman AM, Chee JY, Lakshmanan M, Chin CF, Sudesh K. 2022. Antimicrobial resistance: Prevalence, economic burden, mechanisms of resistance and strategies to overcome. European Journal of Pharmaceutical Sciences. 170. doi:10.1016/j.ejps.2021.106103.

Rahman MRT, Fliss I, Biron E. 2022. Insights in the Development and Uses of Alternatives to Antibiotic Growth Promoters in Poultry and Swine Production. Antibiotics. 11(6). doi:10.3390/antibiotics11060766.

Raza M, Jayasinghe ND, Muslam MMA. 2021. A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms. In: International Conference on Information Networking. Vol. 2021-January. IEEE Computer Society. p. 327–332.

Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T. 2020. Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access. 8:54776–54788. doi:10.1109/ACCESS.2020.2980942.

Riesenfeld CS, Schloss PD, Handelsman J. 2004. Metagenomics: Genomic analysis of microbial communities. Annu Rev Genet. 38:525–552. doi:10.1146/annurev.genet.38.072902.091216.

Ringnér M. 2008. What is principal component analysis? Nat Biotechnol. 26(3):303–304. doi:https://doi.org/10.1038/nbt0308-303.

Rosas M, Bezerra A, Duarte-Neto P. 2013. Use of artificial neural networks in applying methodology for allocating health resources. Public Health Practice. 47(1):128–136. doi:https://doi.org/10.1590/S0034-89102013000100017.

Rossi F, Rizzotti L, Felis GE, Torriani S. 2014. Horizontal gene transfer among microorganisms in food: Current knowledge and future perspectives. Food Microbiol. 42:232–243. doi:10.1016/j.fm.2014.04.004.

Rumelhart DE, Hinton GE, Williams RJ. 1985. Learning internal representations by error propagation. ICS Report 8506. San Diego (CA). [accessed 2023 Jan 25]. https://apps.dtic.mil/sti/pdfs/ADA164453.pdf.

Russell B, Whitehead AN. 1910. Principia Mathematica. 1st ed. Cambridge University Press.

Samreen, Ahmad I, Malak HA, Abulreesh HH. 2021. Environmental antimicrobial resistance and its drivers: a potential threat to public health. J Glob Antimicrob Resist. 27:101–111. doi:10.1016/j.jgar.2021.08.001.

Santoro-Lopes G, Gouvêa EF de. 2014. Multidrug-resistant bacterial infections after liver transplantation: An ever-growing challenge. World J Gastroenterol. 20(20):6201–6210. doi:10.3748/wjg.v20.i20.6201.

Santoso MH. 2021. Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom. Brilliance: Research of Artificial Intelligence. 1(2):54–66. doi:10.47709/brilliance.v1i2.1228.

Saraiva M de MS, Lim K, do Monte DFM, Givisiez PEN, Alves LBR, de Freitas Neto OC, Kariuki S, Júnior AB, de Oliveira CJB, Gebreyes WA. 2022. Antimicrobial resistance in the globalized food chain: a One Health perspective applied to the poultry industry. Brazilian Journal of Microbiology. 53(1):465–486. doi:10.1007/s42770-021-00635-8.

Schijven JF, Blaak H, Schets FM, de Roda Husman AM. 2015. Fate of Extended-Spectrum β-Lactamase-Producing Escherichia coli from Faecal Sources in Surface Water and Probability of Human Exposure through Swimming. Environ Sci Technol. 49(19):11825–11833. doi:10.1021/acs.est.5b01888.

Schonlau M, Zou RY. 2020. The random forest algorithm for statistical learning. Stata Journal. 20(1):3–29. doi:10.1177/1536867X20909688.

Sen PC, Hajra M, Ghosh M. 2020. Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Advances in Intelligent Systems and Computing. Vol. 937. Springer Verlag. p. 99–111.

Sharma A, Kaushik P. 2017. Literature Survey of statistical, deep and reinforcement learning in Natural Language Processing. International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India.:350–354. doi:10.1109/CCAA.2017.8229841.

Sharma M, Sharma S, Singh G. 2018. Performance analysis of statistical and supervised learning techniques in stock data mining. Data (Basel). 3(4). doi:10.3390/data3040054.

Simjee S, McDermott P, Trott DJ, Chuanchuen R. 2018. Present and Future Surveillance of Antimicrobial Resistance in Animals: Principles and Practices. Microbiol Spectr. 6(4). doi:10.1128/microbiolspec.arba-0028-2017.

Sinaga KP, Yang MS. 2020. Unsupervised K-means clustering algorithm. IEEE Access. 8:80716–80727. doi:10.1109/ACCESS.2020.2988796.

Smith C, McGuire B, Huang T, Yang G. 2006. The History of Artificial Intelligence. History of Computing. 27.

Sun S, Wang C, Ding H, Zou Q. 2018. Machine learning and its applications in plant molecular studies. Brief Funct Genomics. 19(1):40–48. doi:10.1093/bfgp/elz036.

Tagliani E, Anthony R, Kohl TA, de Neeling A, Nikolayevskyy V, Ködmön C, Maurer FP, Niemann S, van Soolingen D, van der Werf MJ, et al. 2021. Use of a whole genome sequencingbased approach for Mycobacterium tuberculosis surveillance in Europe in 2017-2019: An ECDC pilot study. European Respiratory Journal. 57(1). doi:10.1183/13993003.02272-2020.

Taylor J, Hafner M, Yerushalmi E, Smith R, Bellasio J, Vardavas R, Bienkowska-Gibbs T, Rubin J. 2014. Estimating the economic costs of antimicrobial resistance: Model and Results. Cambridge, UK. www.rand.org/giving/contribute.

Tiseo K, Huber L, Gilbert M, Robinson TP, van Boeckel TP. 2020. Global trends in antimicrobial use in food animals from 2017 to 2030. Antibiotics. 9(12):1–14. doi:10.3390/antibiotics9120918.

Trotter AJ, Aydin A, Strinden MJ, O'Grady J. 2019. Recent and emerging technologies for the rapid diagnosis of infection and antimicrobial resistance. Curr Opin Microbiol. 51:39–45. doi:10.1016/j.mib.2019.03.001.

Turing AM. 1950. Computing Machinery and Intelligence. Mind. LIX(236):433–460. doi:https://doi.org/10.1093/mind/LIX.236.433.

Uyttendaele M, Jaykus LA, Amoah P, Chiodini A, Cunliffe D, Jacxsens L, Holvoet K, Korsten L, Lau M, McClure P, et al. 2015. Microbial Hazards in Irrigation Water: Standards, Norms, and Testing to Manage Use of Water in Fresh Produce Primary Production. Compr Rev Food Sci Food Saf. 14(4):336–356. doi:10.1111/1541-4337.12133.

Varela MF, Stephen J, Lekshmi M, Ojha M, Wenzel N, Sanford LM, Hernandez AJ, Parvathi A, Kumar SH. 2021. Bacterial resistance to antimicrobial agents. Antibiotics. 10(5). doi:10.3390/antibiotics10050593.

Verraes C, van Boxstael S, van Meervenne E, van Coillie E, Butaye P, Catry B, de Schaetzen MA, van Huffel X, Imberechts H, Dierick K, et al. 2013. Antimicrobial resistance in the food chain: A review. Int J Environ Res Public Health. 10(7):2643–2669. doi:10.3390/ijerph10072643.

Vuttipittayamongkol P, Elyan E, Petrovski A. 2021. On the class overlap problem in imbalanced data classification. Knowl Based Syst. 212. doi:10.1016/j.knosys.2020.106631.

Wang K, Chen Z. 2016. Stepwise Regression and All Possible Subsets Regression in Education. Electronic International Journal of Education, Arts, and Science. 2(Special Issue):60–81. http://www.eijeas.com.

WHO. 2015. Global Action Plan on Antimicrobial Resistance. www.paprika-annecy.com.

WHO, Fund G. 2017. HIV Drug Resistance Report.

Williams KJ. 2009. The introduction of "chemotherapy" using arsphenamine - The first magic bullet. J R Soc Med. 102(8):343–348. doi:10.1258/jrsm.2009.09k036.

World Bank. 2017. Drug-resistant infections: A Threat to Our Economic Future. www.worldbank.org.

Wright GD. 2007. The antibiotic resistome: The nexus of chemical and genetic diversity. Nat Rev Microbiol. 5(3):175–186. doi:10.1038/nrmicro1614.

Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen A, Kjeldsen NJ, Liu C, et al. 2016. A reference gene catalogue of the pig gut microbiome. Nat Microbiol. 1. doi:10.1038/nmicrobiol.2016.161.

Xiong W, Sun Y, Zeng Z. 2018. Antimicrobial use and antimicrobial resistance in food animals. Environmental Science and Pollution Research. 25(19):18377–18384. doi:10.1007/s11356-018-1852-2.

Xiong W, Wang Y, Sun Y, Ma L, Zeng Q, Jiang X, Li A, Zeng Z, Zhang T. 2018. Antibiotic-mediated changes in the fecal microbiome of broiler chickens define the incidence of antibiotic resistance genes. Microbiome. 6(1):1–11. doi:10.1186/s40168-018-0419-2.

Yousefi Milad, Yousefi Moslem, Ferreira RPM, Kim JH, Fogliatto FS. 2018. Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. Artif Intell Med. 84:23–33. doi:10.1016/j.artmed.2017.10.002.

Yuan Y, Wu L, Zhang X. 2021. Gini-Impurity Index Analysis. 16:3154–3169. doi:10.1109/TIFS.2021.3076932.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy. 67(11):2640–2644. doi:10.1093/jac/dks261.

Zhang H, Chen W, Wang J, Xu B, Liu H, Dong Q, Zhang X. 2020. 10-Year Molecular Surveillance of Listeria monocytogenes Using Whole-Genome Sequencing in Shanghai, China, 2009–2019. Front Microbiol. 11. doi:10.3389/fmicb.2020.551020.

Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, Young S, Lam C, Folster JP, Whichard JM, McDermott PF. 2016. Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in Campylobacter spp. Appl Environ Microbiol. 82(2):459–466. doi:10.1128/AEM.02873-15.

Zhou SYD, Wei MY, Giles M, Neilson R, Zheng F, Zhang Q, Zhu YG, Yang XR. 2020. Prevalence of Antibiotic Resistome in Ready-to-Eat Salad. Front Public Health. 8. doi:10.3389/fpubh.2020.00092.

Zhu YG, Zhao Y, Zhu D, Gillings M, Penuelas J, Ok YS, Capon A, Banwart S. 2019. Soil biota, antimicrobial resistance and planetary health. Environ Int. 131. doi:10.1016/j.envint.2019.105059.

Appendix I – R Studio packages

readxl
openxlsx
writexl
dplyr
caret
randomForest
plotly
ggplot2
reshape2
mltools
cluster
qgraph
MASS
cowplot
FactoMineR
factoextra
corrplot
leaps