INSTITUTO
UNIVERSITÁRIO
DE LISBOA

---

mlSoccer: Ferramenta baseada em aprendizagem automática para investigar fatores que influenciam o sucesso do passe

Hugo Vítor Carapinha Muacho

Mestrado em Gestão de Sistemas de Informação

Orientador:
Doutor Rui Jorge Henriques Calado Lopes, Professor Associado
Iscte - Instituto Universitário de Lisboa

Co-Orientador:
Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado
Iscte - Instituto Universitário de Lisboa

Fevereiro, 2023

TECHNOLOGY
AND ARCHITECTURE

Departamento de Ciências e Tecnologias da Informação

mlSoccer: Ferramenta baseada em aprendizagem automática para investigar fatores que influenciam o sucesso do passe

Hugo Vítor Carapinha Muacho

Mestrado em Gestão de Sistemas de Informação

Orientador:
Doutor Rui Jorge Henriques Calado Lopes, Professor Associado
Iscte - Instituto Universitário de Lisboa

Co-Orientador:
Doutor Ricardo Daniel Santos Faro Marques Ribeiro, Professor Associado
Iscte - Instituto Universitário de Lisboa

Fevereiro, 2023

Agradecimentos

Agradeço ao professores Rui Lopes e Ricardo Ribeiro pelos conselhos e conhecimento transmitido na elaboração deste trabalho. Agradeço também ao treinador Nelson Caldeira pelo tempo disponibilizado para a compreensão da base de dados.

Agradeço por todo o apoio à minha família e amigos.

Resumo

Esta dissertação tem como principal objetivo compreender e investigar, através de árvores de decisão e máquinas de vetores de suporte, os fatores que influenciam o sucesso do passe.

A metodologia utilizada é o CRISP DM, usando três iterações do seu ciclo de vida.

Na primeira iteração utilizaram-se as variáveis: área do Voronoi, distância à baliza e variável de saída a continuação da posse de bola. Segundo o especialista consultado, os resultados estavam longe do esperado.

Na segunda iteração é utilizada a mesma base de dados, considerando o penúltimo e último evento (passe) da jogada. Nesta fase não foram encontradas ligações entre a distância à baliza e área total do Voronoi no sucesso ou não do passe.

Na última iteração são realizadas três experiências com diferentes conjuntos de dados obtidos da mesma base de dados: um contendo dados posicionais e área de Voronoi dos jogadores, outro contendo somente os dados do jogador que executa o passe, o terceiro considerando o portador da bola e os companheiros e adversários mais próximos. Os resultados obtidos do primeiro conjunto de dados indicam que a importância relativa das variáveis é dependente do jogo e de alguma forma relacionada com a formação das equipas e missão tática dos jogadores. No segundo conjunto de variáveis os resultados indicam que as variáveis mais importantes são a área do Voronoi e a distância à baliza sendo as coordenadas dos jogadores pouco relevantes. O terceiro conjunto, mais diretamente relacionado ao processo de aprovação, proporcionou uma classificação mais consistente das características. Valores baixos da medida F e precisão mostram que as variáveis e os fatores que levam ao sucesso da aprovação são de fato elusivos.

Palavras-chave: Aprendizagem Automática, Árvores de Decisão, Máquinas de Vetor de Suporte, Futebol, Análise de Desempenho, Sucesso do Passe

Abstract

This dissertation has as main objective to understand and investigate, through decision trees and support vector machines, the factors that influence the success of the pass.

The methodology used is CRISP DM, using three iterations of its life cycle.

In the first iteration the following variables were used: Voronoi area, distance to goal and output variable the continuation of possession. According to the expert consulted, the results were far from what was expected.

In the second iteration the same database is used, considering the penultimate and last event (pass) of the play. At this stage no links were found between the distance to the goal and total area of the Voronoi in the success or not of the pass.

In the last iteration three experiments are performed with different data sets from the same database: one containing positional data and Voronoi area of all players, another containing only the data of the player performing the pass, the third considering the ball carrier and the closest teammates and opponents. The results obtained from the first set of data indicate that the relative importance of the variables is game dependent and somewhat related to the formation of the teams and tactical mission of the players. In the second set of variables the results indicate that the most important variables are the Voronoi area and the distance to the goal with the players' coordinates being of little relevance. The third set, more directly related to the approval process, provided a more consistent classification of characteristics. Low values of the F-measure and hit show that the features and factors that lead to successful passing are indeed elusive.

Keywords: Machine Learning, Decision Trees, Support Vetor Machine, soccer, performance analysis, pass success

Conteúdo

Lista de Figuras

Lista de Tabelas

CAPíTULO 1

Introdução

Vivemos atualmente na *Era Digital* [1], e quase diariamente que vemos novas tecnologias a surgir e a ter o seu lugar na sociedade, sendo a Inteligência Artificial (IA) uma tecnologia que tem acompanhado essa contribuição para a sociedade, tornando-se assim numa da maiores tendências no mundo atual. A recente popularidade da IA pode ser atribuída aos seguintes três fatores: o crescimento do *Big Data*, o acesso fácil a recursos computacionais e o desenvolvimento de novas técnicas de IA. A IA surge, para o mundo dos negócios, como uma possível solução para lidar com as grandes quantidades de dados com que as empresas se deparam atualmente.

O futebol não foge à regra da sociedade e tem acompanhado o crescimento da IA, através da introdução da mesma em diversas áreas. Um exemplo de utilização é a avaliação do desempenho do jogador através da localização geográfica de todos os jogadores em campo melhorando a coordenação da equipa. Outro exemplo é também o *expected goals* que de forma simples, diz-nos a quantidade, mas principalmente a qualidade de oportunidades de golo que uma equipa criou num jogo. Com a introdução da IA no futebol as equipas são capazes de descobrir novos potenciais e atingir novos e ambiciosos objetivos, especialmente no aumento da competitividade da equipa e melhorar a tomada de decisão Apesar disso a tecnologia ainda é imatura e precisa de significativas melhorias [2].

1.1. Motivação

A tecnologia e a IA influenciam cada vez mais o nosso quotidiano, passando de uma pequena ajuda para um crescimento disruptivo, transformando-se assim em algo muito mais poderoso do que alguma vez testemunhámos. Este crescimento está bem visível no aumento das tecnologias utilizadas no futebol, como diz Esteban Granero, ex-jogador de futebol e líder uma empresa de consultora especializada em futebol.“A IA permite

que os clubes usem análises preditivas e prescritivas para reduzir as incertezas e tomar melhores decisões".

1.2. Questão de Investigação e Objetivos

Hoje em dia a tecnologia já é muito utilizada no futebol nomeadamente no processo de análise de desempenho. Neste processo existem dois passos muito importantes: a recolha de dados e o processamento dos dados. Em relação ao primeiro passo, existem dois grandes grupos. Um grupo que tem como base a utilização de aplicações informáticas para anotação e outro que tem como base de desenvolvimento a obtenção de dados de forma automática.

O objetivo desta dissertação está focado no segundo passo, ou seja, na análise dos dados. A análise de dados pode ser feita de várias formas, por exemplo através de indicadores e estatísticas que neste momento é a forma que tem sido mais adota. Esta dissertação, para além dessa análise de indicadores, tem como objetivo geral a criação de uma ferramenta baseada em IA que ajude elementos das equipas técnicas a compreender os fatores que influenciam o resultado do gesto técnico passe.

Esta temática, centrada na IA ligada ao futebol, leva-nos à seguinte questão de investigação:

- Como é que uma técnica de IA pode ser usada para compreender os fatores que influenciam o sucesso de um gesto técnico?

Os objetivos específicos desta dissertação são os seguintes:

- Compreender os conceitos fundamentais no futebol que são usados para a compreensão da dinâmica das equipas (e.g., sinergia) e gestos técnicos dos jogadores (e.g., passe);
- Estudo de diferentes ferramentas de aprendizagem automática e avaliação da sua aplicabilidade à compreensão do gesto técnico passe (sucesso/insucesso);
- Identificação dos fatores significativos, em diferentes ferramentas, para a compreensão do gesto técnico passe (e.g., relação entre a posição dos jogadores e o evento que aconteceu na jogada).

Para avaliar o cumprimento destes objetivos ir ão ser utilizadas duas abordagens: uma avaliação objetiva recorrendo a uma modelação e avaliação das técnicas utilizadas e outra abordagem mais subjetiva através de perguntas a especialistas sobre a utilidade e eficácia da ferramenta.

A principal contribuição desta dissertação será a obtenção de uma ferramenta de IA para compreender as principais variáveis que influenciam o passe. Com a validação desta ferramenta, espera-se aproximar mais e melhor os intervenientes da modalidade à IA.

1.3. Metodologia

No presente trabalho foram consideradas duas metodologias para abordar a questão de investigação, são elas: *Development Research* (DR) e CRISP DM.

A DR tem como objetivo de pesquisa estabelecer uma base empírica para a criação de produto e modelos novos ou aprimorados que governam o desenvolvimento [3].

Quanto ao CRISP DM tem como utilidade transformar dados de uma empresa ou de um determinado *dataset* em informação e conhecimento úteis [4]. Esta metodologia é composta por seis fases e poderá ter várias iterações ao longo do processo antes de avançar para as próximas fases sempre que algo não pareça correto e necessite de ajustes.

Depois de analisar as duas metodologias foi decidido que a melhor abordagem ao problema seria através do CRISP DM. Isto porque, o problema em questão centra-se muito na leitura e transformação dos dados através de técnicas de modelação com o objetivo de extrair informação útil (neste caso particular a compreensão dos fatores de sucesso do gesto passe). Na Figura 1 podemos ver todas as etapas da metodologia.

A primeira etapa do CRISP DM "Compreensão do Problema" consiste em saber e delinear qual o objetivo da investigação e fazer uma revisão da literatura de todos os temas abordados na investigação. Na dissertação, nesta etapa foi definida a questão de investigação, a motivação, os objetivos e também abordados temas como a sinergia, IA e aprendizagem automática.

Na segunda etapa, "Compreensão dos Dados", direciona o foco para identificar e analisar os conjuntos de dados que ajudam a cumprir os objetivos do projeto. Nesta

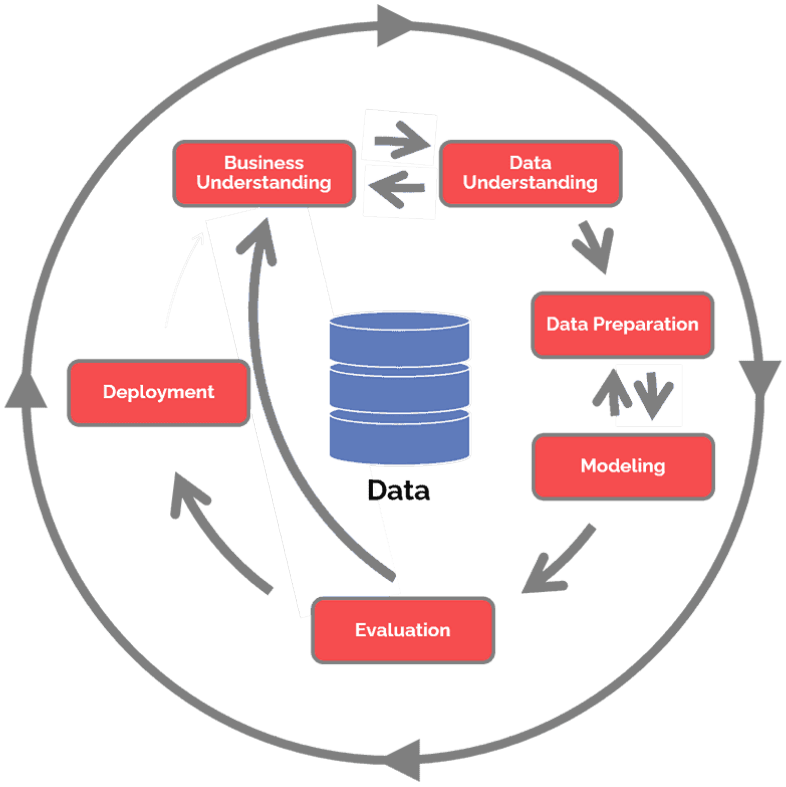

Figura 1. Diagrama CRISP DM (retirada de `https://www.datascience-pm.com/crisp-dm-2/`)

etapa foi feita uma análise minuciosa de todas as variáveis do conjunto de dados e verificou-se se existiam valores omissos ou caracteres inválidos.

A etapa seguinte é a "Preparação dos Dados" que consiste em organizar a informação para utilizar mais tarde. É feita uma limpeza dos dados e uma seleção das variáveis mais importantes para responderem à questão de investigação.

Depois de feita a preparação dos dados foi feita uma análise a várias técnicas de aprendizagem supervisionada e de seguida uma seleção de qual aplicar para extrair o conhecimento, a esta etapa chama-se "Modelação".

Após a aplicação de várias técnicas é escolhida aquela que melhor representou e respondeu aos objetivos iniciais, a isto chama-se "Avaliação".

1.4. Estrutura do Documento

A presente dissertação encontra-se organizada em quatro capítulos. É de seguida apresentado um breve resumo de cada capitulo.

Após o capítulo inicial, a Introdução, o segundo capítulo faz um enquadramento conceptual, em que se abordam as definições de temas importantes relativos à dissertação, como o gesto técnico passe, sinergia coletiva e também da aprendizagem automática (supervisionada e não supervisionada) e como eles se relacionam. É também feito uma análise aos trabalho já publicados sobre estes dois conceitos. O conteúdo do capítulo 2 resulta da aplicação da primeira etapa da metodologia CRISP DM.

No terceiro capitulo são abordadas e analisadas as experiências feitas através da metodologia CRISP DM. São também analisados os resultados obtidos em cada uma das iterações da metodologia. O capítulo 3 subdivide-se em três secções que correspondem à aplicação em três iterações dos passos dois a quatro da metodologia. No quinto capítulo são apresentadas as conclusões da dissertação.

CAPíTULO 2

Enquadramento Conceptual

Como referido anteriormente, um dos objetivos desta dissertação é a criação de uma ferramenta baseada em IA que ajude elementos das equipas técnicas a compreender o resultado do gesto técnico passe. Para compreender esses resultados é preciso perceber primeiramente alguns conceitos fundamentais da dinâmica do futebol, como compreender o gesto técnico a avaliar (passe) e também a sinergia.

2.1. Gesto Técnico Passe

Nesta secção irá ser abordado o gesto técnico passe através da sua definição e da sua importância num jogo de futebol. O passe vai ser abordado em termos do seu sucesso e não como é executado em termos corporais do executante.

Também nesta secção, iremos analisar e comentar alguns artigos relacionados com o passe de modo a perceber o que tem sido explorado neste tema.

2.1.1. Definição e Importância

O passe pode ser definido, muito resumidamente, por passar a bola a um companheiro de equipa [5]. O passe é um dos muitos gestos técnicos e ações no futebol e o mais utilizado e importante para avançar com sucesso a bola no campo de modo a chegar ao objetivo final, o golo [6]. Para chegar a esse objetivo as equipas manipulam o espaço para condicionar o passe. Para o executante é importante saber para onde e o porquê de executar o passe de modo a ter sucesso. Durante um jogo existem um número significativo de passes (na ordem das centenas) e alguns com mais importância que os outros. Os passes que são efetuados entre dois defesas-centrais no seu próprio meio-campo e um passe efetuado pelos jogadores mais avançados no terreno têm uma importância e

uma probabilidade de errar muito diferente. O passe efetuado pelos atacantes no meio-campo ofensivo é claramente mais valioso e pode criar uma oportunidade de golo e tem uma probabilidade de errar maior.

Os passes efetuados por uma equipa podem também servir para distinguir a abordagem tática das equipas. Enquanto umas equipas optam por criar oportunidades de golo através de passes mais longos e diretos, outras optam por um jogo mais elaborado e de posse [7]. A principal fraqueza das abordagens atuais para avaliar o passe e o comportamento tático em geral é que elas raramente incluem variáveis contextuais e a interação com o oponente [8]. Várias abordagens foram desenvolvidas para medir a qualidade de um passe tendo em conta a oposição. A avaliação era somente baseada na probabilidade da equipa marcar golo. Ou seja, um passe era classificado como "bom", se o mesmo produzisse uma oportunidade, ou aumentasse a probabilidade dentro de um determinado intervalo de tempo [9]. Para avaliar a eficácia do passe, Tenga, Ronglan e Bahr [10] dividiram os passes em duas categorias: penetrativo ou não penetrativo. O passe penetrativo seria quando a equipa atacante consegue uma vantagem posicional sobre a equipa adversária. Entenda-se como vantagem posicional quando a equipa atacante através do passe estabelece uma situação de maioria na área adversária. O passe não penetrativo a equipa que executa o passe não obtém nenhuma vantagem posicional com o passe.

2.1.2. Trabalho Relacionado

Nesta secção irão ser apresentados exemplos de trabalhos realizados na área do gesto técnico passe. Os passes podem ser estudados de várias maneiras: no seu processo e avaliação (área em que esta dissertação se concentra) ou no processo morfológico e corporal da execução do passe. Irão ser apresentados estudos nestas duas áreas.

Rein, Raabe e Memmert [9] apresentam duas abordagens para avaliar a eficácia do passe no futebol, muito semelhantes ao que será feito nesta dissertação. Uma das abordagens pretende avaliar o sucesso em situações de maioria, isto é, calculando o número de defensores entre o portador da bola e o golo. A outra abordagem é consoante o

controlo do espaço que é estimado através de diagramas de Voronoi com base nos jogadores em campo. Os resultados deste estudo mostram que ambas as medidas estão significativamente relacionadas ao sucesso do jogo no que diz respeito ao número de golos marcados e à probabilidade de ganhar um jogo.

Também próximo do trabalho apresentado nesta dissertação, Arbués Sangüesa, Martin, Fernandez et al. [11] estudaram o uso de um modelo computacional para estimar a viabilidade de um passe. O modelo utiliza como variáveis a orientação do jogador que executa o passe e a configuração espacial dos oponentes. Depois de analisados mais de 6000 eventos de passes, o estudo concluiu que incluindo a orientação como medida para calcular a viabilidade o modelo construido atingiu mais de 0,70 de taxa de acerto.

Por último e numa área completamente diferente da desta dissertação, Keskin [12] visa compreender os efeitos do aquecimento com bolas de dois tamanhos diferentes. O estudo foi avaliado através de um método que passa por completar 16 passes com sucesso no menor tempo possível. Neste estudo eram divididos os participantes em dois grupos de 14 pessoas. O primeiro grupo aquecia com a mesma bola com que fazia o teste, enquanto o outro grupo aquecia com uma bola menor do que com aquela que fazia o teste. O estudo concluiu que as habilidades de passe foram afetadas positivamente durante a partida entre os jogadores de futebol que aqueceram com bolas de tamanho menor antes da partida.

2.2. Sinergia em Desportos Coletivos

Nesta secção irá ser abordada a definição de sinergia ligada aos desportos coletivos, sendo explicados alguns exemplos da aplicação deste conceito nesses desportos. O objetivo de compreender este conceito é perceber e olhar o gesto técnico passe como um ato e uma ação conjunta e não como dependente exclusivamente do seu executante.

Segunda parte desta secção vão ser explicadas as características principais para considerarmos que existe sinergia e qual a forma de quantificar essa mesma sinergia.

Na terceira parte da secção, que é o "Trabalho Relacionado" vai ser feita uma comparação de alguns artigos que abordam a sinergia no desporto coletivo principalmente ligados ao futebol com a dissertação que vai ser desenvolvida.

2.2.1. Definição

Sinergia é definida como esforço ou ato coletivo na ação conjunta de vários elementos ou de várias partes que pretende obter um resultado melhor ou maior do que a soma das partes.[1] Sendo o passe também uma ação conjunta, é plausível que existam sinergia durante o mesmo. As sinergias são essencialmente emergentes, i.e., não são concebidas por *design*, nem são programadas para ocorrerem de uma forma pré-organizada. Numa sinergia de grupo, como é o caso do desporto coletivo, as decisões e ações dos jogadores para formar uma sinergia não devem ser vistas como independentes, assim os vários jogadores sincronizam atividades de acordo com ambientes dinâmicos de desempenho em frações de segundo [13]. Para o desporto coletivo uma sinergia é uma tarefa coletiva de uma organização de indivíduos com funções específicas, de modo que os graus de liberdade de cada indivíduo no sistema sejam acoplados, permitindo assim que esses mesmo graus de liberdade de diferentes indivíduos se co-regulem uns aos outros [14]. As sinergias requerem a modulação de menos parâmetros do que o controle separado de cada grau de liberdade, a fim de produzir um movimento coordenado. Essa tarefa do sistema reduz a necessidade de controle de cada grau de liberdade e permite que uma ação de um elemento da sinergia seja compensado por outro. Uma característica importante de uma sinergia coletiva é a capacidade de um indivíduo (por exemplo, um jogador de uma equipa) para influenciar comportamentos de outros.

Definindo tarefa defensiva como um conjunto de ações que os jogadores executam para contrariar a equipa atacante de chegar à baliza e marcar golo, podemos dizer que no futebol a sinergia é utilizada por exemplo numa tarefa defensiva [15]. Por exemplo no basquetebol, um defensor ao ter informação sobre a mão dominante do opositor e a posição relativa do cesto restringe as posições e as tendências de coordenação do 1 *vs* 1, atacante-defensor [16].

Usando dados posicionais obtidos do rastreamento dos jogadores, estudos recentes mostram como jogadores da mesma equipa interagem continuamente durante uma jogada. Por exemplo, foi observado que as equipas tendem a ser perfeitamente sincronizados nos seus movimentos laterais e longitudinais [14]. Os padrões de coordenação

[1] https://dicionario.priberam.org/sinergia

observados mostraram comportamentos compensatórios dentro da equipa, estes comportamentos são característicos de uma sinergia [17].

Como foi referido anteriormente, esta dissertação visa olhar para o passe como uma tarefa coletiva e potencialmente influenciada por todos os jogadores, em particular por quem executa o passe. Na execução do passe existem dois conjuntos de jogadores com tarefas diferentes, a equipa com posse de bola que tem como objetivo o sucesso do passe e a equipa adversária que pretende que o passe não tenha sucesso.

2.2.2. Características e Quantificação

Como referido as sinergias requerem a modulação de menos parâmetros do que o controlo separado de cada grau de liberdade, a fim de produzir um movimento coordenado. Para que um grupo de componentes seja considerado uma sinergia têm que estar presentes as seguintes características:

- todos os intervenientes devem contribuir para o desempenho de uma determinada tarefa;
- compensação de erros, alguns elementos podem apresentar alterações na sua tarefa para compensar um elemento que pode não estar a fazer uma tarefa relevante;
- dependência de tarefa, formar uma sinergia diferente para um propósito diferente com base no mesmo conjunto de componentes.

No caso do gesto técnico passe verificamos estas características porque:

- os intervenientes de um jogo quando é executado o passe têm um objetivo: a equipa com posse de bola o sucesso do passe e a equipa adversária o contrário. Todos os intervenientes executam uma tarefa para conseguir chegar a esse objetivo;
- As tarefas dos jogadores podem mudar consoante o posicionamento dos seus colegas de equipa, através de uma compensação no seu posicionamento ou numa desmarcação por exemplo.

A medição da sinergia é feita por meio da medição das principais tarefas do sistema, considerando a contribuição de cada indivíduo [14]. Estas incluem:

- compreensão dimensional, um processo do qual resulta um conjunto de graus independentes de liberdade, no qual a sinergia surge;
- compensação recíproca, se um elemento não produz a sua função, outros elementos devem exibir mudanças nas suas contribuições (ações) para que os objetivos da tarefa continuem a ser atingidos;
- ligações interpessoais, a contribuição específica de cada elemento para uma tarefa de grupo.

Um exemplo de como medir uma sinergia coletiva é através do método de *Uncontrolled Manifold Hypothesis* (UCM). Este método é usado para quantificar sinergias motoras, definidas como uma organização específica do sistema nervoso central que mantém a estabilidade das ações motoras específicas da tarefa, o UCM permite a análise de variância entre tentativas consecutivas.

Nesta dissertação não usamos diretamente nenhuma destas medidas para calcular a sinergia. No entanto, a abordagem adotada está relacionada com a compreensão dimensional, uma vez que se tenta identificar um conjunto reduzido de variáveis que têm impacto no sucesso do passe. Também as ligações interpessoais estão presentes no contexto deste trabalho: por exemplo ao tentar perceber a importância dos jogadores mais próximos do jogador que executa o passe.

2.2.3. Trabalho Relacionado

Nesta secção serão apresentados alguns exemplos de trabalhos realizados sobre sinergia ligados ao desporto, nomeadamente o futebol.

O estudo de Milho e Passos [15] tinha como objetivo quantificar as sinergias coletivas no futebol, em particular, quantificar as sinergias interpessoais entre os quatro defesas que formavam a linha defensiva. A hipótese deste estudo era que os jogadores que formavam a linha defensiva ajustavam as suas posições relativas para estabilizar uma distância interpessoal e, como tal, criar uma sinergia coletiva. Para identificar as sinergias que eram formadas pelos defesas durante o ataque foi usado o UCM. Dos resultados obtidos neste estudo verificou-se que das dezoito situações estudadas, em apenas três delas não foram identificadas a formação de sinergia. Isto sugere que os movimentos

dos defensores contribuem para estabilizar as distâncias interpessoais, formando assim a sinergia coletiva.

Já Esteves, Araujo, Davids et al. [16] estudaram no basquetebol os efeitos do posicionamento relativo das díades atacante-defensor ao cesto sobre as tendências de coordenação interpessoal. As trajetórias de deslocamento do movimento dos jogadores foram gravadas em vídeo e digitalizadas em 162 tentativas. Os resultados mostraram que as tendências de coordenação interpessoal mudaram de acordo com a distância da posição relativa dos jogadores ao cesto. Os resultados também demonstraram como a informação sobre a mão dominante do atacante e a posição relativa ao cesto podem restringir as tendências de coordenação do atacante-defensor nas sub fases 1 *vs* 1.

Num outro estudo [17], foi proposto um método espaço-temporal qualitativo baseado no QTC (Cálculo de Trajetória Qualitativa) para análise de formação de equipas no futebol. O Cálculo de Trajetória Qualitativa permite a comparação das formações das equipas ao descrever as relações relativas entre os jogadores de maneira qualitativa. Com este método, as formações das equipas, tanto no seu conjunto total de jogadores como só um grupo menor de jogadores podem ser descritas. A análise dessas formações pode ser feita para várias partidas, definindo assim o estilo de jogo de uma equipa. Além disso, diferentes pesos podem ser alocados para os vetores entre os jogadores, de acordo com a sua importância no campo.

Com a análise deste três estudos podemos concluir que já tem sido estudado a sinergia nos desportos coletivos e que o conceito de sinergia tem muito impacto para explicar dinâmicas da equipa em todos os seus processos táticos.

2.3. Aprendizagem Automática

Como já foi referido um dos objetivos desta dissertação é criar uma ferramenta baseada em IA que ajude a compreender os fatores que influenciam o sucesso do gesto técnico passe.

Para responder ao objetivo é preciso primeiro compreender o conceito de aprendizagem automática, nas suas duas vertentes (supervisionada e não supervisionada), perceber quais as técnicas que poderão ser utilizadas e as suas vantagens e desvantagens e por último ligar estes conceitos referidos anteriormente ao futebol, principalmente na vertente tática.

Na primeira secção deste capítulo irá ser abordado o conceito de aprendizagem automática através da sua definição.

Na segunda secção irão ser abordados o conceito de aprendizagem supervisionada, alguns exemplos da sua utilização e objetivos, bem como serão analisadas duas técnicas de aprendizagem supervisionada (máquinas de vetor de suporte e árvores de decisão).

Na terceira secção será aprofundado o conceito de aprendizagem não supervisionada, os seus objetivos e técnicas que são utilizadas neste tipos de métodos.

Por último, na quarta secção vai ser feita uma análise a técnicas de aprendizagem automática e de IA já utilizadas no futebol nas diversas áreas e fazer uma ligação destas técnicas com a dissertação.

2.3.1. Definição

A Aprendizagem Automática é o campo de estudo que permite que computadores aprendam sem que sejam programados. Pode ser definida como um conjunto de métodos que podem detetar automaticamente padrões nos dados e usar esses padrões para prever dados futuros ou para realizar outros tipos de tomada de decisão [18]. A aprendizagem automática está a começar a desempenhar um papel essencial dentro dos seguintes ramos da computação: migração de dados, aplicativos difíceis de programar e aplicativos de software personalizados [19]. Os algoritmos de aprendizagem automática geralmente enquadram-se em dois paradigmas: aprendizagem supervisionada e aprendizagem não supervisionada [20].

As empresas de análise de futebol só recentemente começaram a analisar os chamados *big data* (por exemplo, vídeos de alta resolução, rastreamento do movimento dos jogadores e informações de posse). A quantidade de dados disponíveis no futebol aumentou com diferentes técnicas para recolher uma grande quantidade de dados como por exemplo: sensores, GPS e algoritmos de visão computacional. Ao mesmo tempo, apenas recentemente grandes avanços foram feitos na aprendizagem automática, produzindo técnicas que podem lidar com esses novos conjuntos de dados de alta dimensão. Isto ajuda a utilização da aprendizagem automática no futebol nas suas várias áreas como por exemplo: no recrutamento e desempenho de jogadores, na venda de bilhetes de modo a aproximar os adeptos ao seu clube e também na ajuda de tomadas de decisão que atinjam toda uma área de um clube.

2.3.2. Aprendizagem Supervisionada

Na aprendizagem supervisionada assume-se a presença de um "professor", onde é fornecida informação etiquetada ("respostas corretas") para cada situação. As técnicas de aprendizagem supervisionada constroem modelos preditivos que aprendem a partir de um grande número de exemplos de treino, onde cada exemplo de treino tem um rótulo que indica a sua saída correta [21]. Na aprendizagem supervisionada existe um par que consiste no objeto de entrada e um valor rótulo de saída pertencente a uma classe ou a valores contínuos. Em geral distinguem-se os seguintes tipos de aprendizagem supervisionada:

- Classificação, saídas com valores discretos (classes);
- Regressão, saídas com valores numéricos com uma redução de ordem, como valores contínuos, reais.

2.3.2.1. *Máquinas de Vetores de Suporte*

Uma máquina de Suporte Vetorial é um algoritmo de aprendizagem automática usado tanto para problemas de classificação como regressão. Uma das vantagens deste método é o melhor desempenho com um número limitado de amostras [22]. O classificador tem como objetivo encontrar uma função que permita classificar os dados corretamente, ou

pelo menos da melhor forma possível, evitando ajustar-se demasiado ao conjunto de treino dos dados [23].

Os principais conceitos associados a este algoritmo são os seguintes:

hiperplano: é um plano de decisão que separa um conjunto de objetos diferentes;
vetores de suporte: são os pontos de dados mais próximos do hiperplano;
margem: é a separação entre os hiperplanos paralelos ao hiperplano que contém os pontos mais próximos das classes diferentes. Se a margem for maior entre as classes, então é considerada uma margem boa.

Com as máquinas de vetores de suporte pretende-se conseguir o melhor classificador possível que corresponderá àquele que apresenta menor risco empírico e satisfaça as respetivas restrições.

Este algoritmo é implementado através da utilização de um *kernel*. Um kernel transforma um espaço de dados de entrada no num outro espaço onde um hiperplano possa ser usado. O algoritmo tenta converter problemas não separáveis em problemas separáveis. Alguns tipos de *kernel* são os seguintes:

- linear: pode ser usado como produto escalar normal em qualquer observação dada. O produto entre os dois vetores é a soma da multiplicação de cada par de valores de entrada;
- polinomial: é uma forma mais generalizada do kernel linear. O kernel polinomial permite distinguir objetos de classes diferentes num espaço onde as classes não são linearmente separáveis, mas podem ser por um polinómio;
- radial: é uma função de kernel popular usada normalmente na classificação de máquinas de vetor de suporte baseada na função exponencial. Pode mapear um espaço de entrada em espaço dimensional infinito (ou, pelo menos tão grande como o conjunto de dados);

2.3.2.2. *Árvores de Decisão*

As árvores de decisão são um algoritmo de aprendizagem supervisionada, utilizada em tarefas de classificação e regressão. Isto é, pode ser usado para prever categorias discretas (sim ou não, por exemplo) e para prever valores numéricos (o valor do lucro em

reais). O objetivo é criar um modelo que preveja o valor de uma variável de destino, aprendendo regras de decisão simples inferidas a partir de dados [22]. São representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados. Como a árvore de decisão segue uma abordagem supervisionada, o algoritmo é alimentado com uma coleção de dados pré-processados. Esses dados são usados para treinar o algoritmo.

Na árvore de decisão o nível superior chama-se raiz, a raiz dá origem a ligações a outros elementos chamados nós. Nós que não tiverem ligações são chamados nós terminais. Os nós de decisão ou simplesmente nós da árvore são as questões que são apresentadas pela árvore depois de passar por cada nó (começando pelo nó raiz). Cada aresta da árvore corresponde ao resultado da questão e o resultado é representado por um nó folha ou um nó terminal que representa a distribuição da classe. Este método tem por base algoritmos que dividem o conjunto inicial de dados em subconjuntos mais homogéneos que por sua vez se podem dividir em subconjuntos ainda mais homogéneos [24]. O algoritmo da árvore de decisão funciona através de várias instruções *if-else* alinhadas em que condições sucessivas são verificadas, a menos que o modelo chegue a uma conclusão.

A principal conjuntura da dissertação é que a importância de uma variável (por exemplo, coordenada x) pode ser quantificada através da sua importância na árvore de decisão. A importância da variável é calculada com a divisão da diminuição da impureza do nó ponderada pela probabilidade de atingir esse nó (o número de amostras que atingem o nó, dividido pelo número total de amostras). A importância de cada nó da árvore de decisão é calculada então usando a Eq. 2.1 [25].

$$\sigma_j = w_j C_j - w_{\mathit{left}(j)} C_{\mathit{left}(j)} - w_{\mathit{right}(j)} C_{\mathit{right}(j)} \tag{2.1}$$

σ_j	importância do nó j
w_j	probabilidade de alcançar o nó j
C_j	o valor da impureza do nó j
$\mathit{left}(j)$	nó filho da esquerda dividido em nój
$\mathit{right}(j)$	nó de filho da direita dividido em nó j

2.3.3. Aprendizagem não Supervisionada

A aprendizagem não supervisionada usa dados de treino não anotados, não classificados e categorizados. O principal objetivo da aprendizagem não supervisionada é descobrir padrões ocultos e interessantes em dados não rotulados [26]. Na aprendizagem não supervisionada não é fornecida nenhuma indicação externa e a aprendizagem é realizada pela descoberta de regularidades (semelhanças) nos dados de entrada, procurando agrupamentos (*clustering*) nos exemplos de treino. Usualmente, o investigador nem sabe quantas classes ou componentes discriminatórias vão ser produzidas após a utilização do algoritmo não supervisionado. Como exemplos de aprendizagem não supervisionada podem ser destacados a descoberta de *clusters* e os modelos de variáveis latentes.

As técnicas de aprendizagem não supervisionada são as seguintes:

- *clustering*, que permite dividir automaticamente o conjunto de dados em grupos de acordo com uma função de similaridade;
- deteção de anomalias pode descobrir automaticamente pontos de dados incomuns em um conjunto de dados;
- associação: identifica conjuntos de itens que frequentemente ocorrem juntos em seu conjunto de dados.

2.3.4. Trabalho Relacionado

A aprendizagem automática e a IA têm sido cada vez mais utilizadas no mundo do futebol, não só no domínio do desempenho ou análise tática mas também noutras áreas tais como no departamento médico e para aproximar os adeptos aos seus clubes.

Um exemplo da aplicação da aprendizagem automática fora do ramo tático é a prevenção de lesões, exemplo disso é um estudo realizado por Rommers, Rössler, Verhagen et al. [27] que durante uma época tentou prever as lesões de 734 jogadores com idades compreendidas entre os 10 e 15 anos de idade de 7 academias belgas. No início da época foi realizado uma bateria de exames para avaliar a coordenação motora e aptidão física, também foi tido em conta a altura, peso força, flexibilidade entre outras características físicas. Com base nessas características, o algoritmo de aprendizagem automática utilizado conseguiu prever lesões e distinguir lesões graves de leves com uma elevada taxa

de acerto. A aplicação deste tipo de algoritmos também ajuda os treinadores na tomada de decisão durante o jogo, como por exemplo saber a condição física de um jogador e se deve ser substituído.

Outro exemplo da aplicação da aprendizagem automática no futebol é na análise de desempenho dos jogadores. Um estudo de Jamil, Phatak, Mehta et al. [28] aplicou vários algoritmos de aprendizagem automática para classificar o desempenho dos guarda-redes profissionais com o objetivo de distinguir um guarda-redes de elite de um guarda-redes sub-elite. Os algoritmos utilizados foram os seguintes: Regressão Logística, *Gradient Boosting Classifiers* e *Random Forest Classifiers*. As conclusões tiradas neste estudo é que todos os guarda-redes de elite tinham as mesmas características comuns: distribuição curta, passar a bola com sucesso, receber passes com sucesso, e não sofrer golos. Em última análise, os resultados descobertos neste estudo sugerem que a habilidade de um guarda-redes com os pés e não necessariamente com as mãos é o que distingue os guarda-redes de elite da sub-elite. Nesta área da análise de desempenho outro estudo [29] também propõe uma plataforma orientada em dados que oferece uma avaliação multidimensional e baseada em princípios do desempenho dos jogadores. Este tipo de plataforma pode ser utilizado na observação de jogadores por parte dos olheiros ou analistas.

Outra aplicabilidade da aprendizagem automática no mundo do futebol é através da análise tática, quer para prever táticas e movimentações dos adversários quer analisar os movimentos e condição física. Exemplo desta aplicabilidade é um modelo baseado em dados que avalia as ações dos jogadores de futebol em relação à sua contribuição para as fases em que a equipa tem posse de bola. Este modelo consiste em prever os movimentos dos jogadores e prever também os resultados de uma posse de bola. As interações entre jogadores e uma bola são capturadas através de uma rede neuronal e os resultados são bastantes positivos tanto ao prever de forma confiável as trajetórias dos jogadores quanto ao resultado das posse de bola (posse de bola perdida ou ganha).

Depois da análise de todos estes estudos podemos confirmar que a aprendizagem automática já é utilizada em muitos ramos no futebol mundial. Esta dissertação está

mais ligada ao ramo da análise tática e análise de dados através da análise de indicadores que ajudam a perceber o sucesso do gesto técnico passe.

2.4. Dados do Jogo e Anotação

As anotações no futebol são uma ferramenta importante para obter dados de um jogo. A análise de partidas de futebol baseia-se na anotação das ações individuais dos jogadores (por exemplo, passes e remates), desempenho físico e ações da equipa (por exemplo, substituições). Consequentemente, anotar eventos de futebol a um nível detalhado é uma tarefa muito cara e propícia a erros [30].

Os dados posicionais geralmente são obtidos por meio de ferramentas automatizadas ou semiautomáticas que contam com dispositivos como recetores GPS, câmaras e visão computacional. Uma das oportunidades mais interessantes proporcionadas pela disponibilidade de dados através do rastreamento de posição no futebol é a análise do comportamento tático. O comportamento tático é um dado importante do desempenho nos desportos coletivos como o futebol, e refere-se a como uma equipa gere o seu posicionamento espacial ao longo do tempo para atingir um objetivo comum. No futebol a escolha da tática certa pode ser determinante no resultado final da partida, podemos definir a tática como a maneira de a equipa gerir o espaço, tempo e ações individuais durante um jogo [31]. A escolha da tática é feita consoante as anotações e análises detalhadas para encontrar comportamentos e padrões na equipa adversária. No futebol moderno de hoje, as equipas jogam em formações táticas específicas. A formação mais comum é o 4-4-2, o que significa quatro defensores em linha reta paralelamente à linha da baliza, quatro meio-campistas e na frente dois atacantes. No entanto, isso é resultado de uma longa evolução. Por exemplo no inicio do futebol, antes da regra do fora de jogo, a tática mais utilizada era 2-3-5.

Um exemplo de um dado do jogo e que vai ser utilizado nesta dissertação é o diagrama de Voronoi. Um diagrama de Voronoi é uma partição de um plano em regiões próximas a cada um de um determinado conjunto de objetos. No caso mais simples, esses objetos são pontos no plano (chamados de sementes, sítios ou geradores). A cada objeto corresponde uma região definida pelos pontos do plano que estão mais próximos

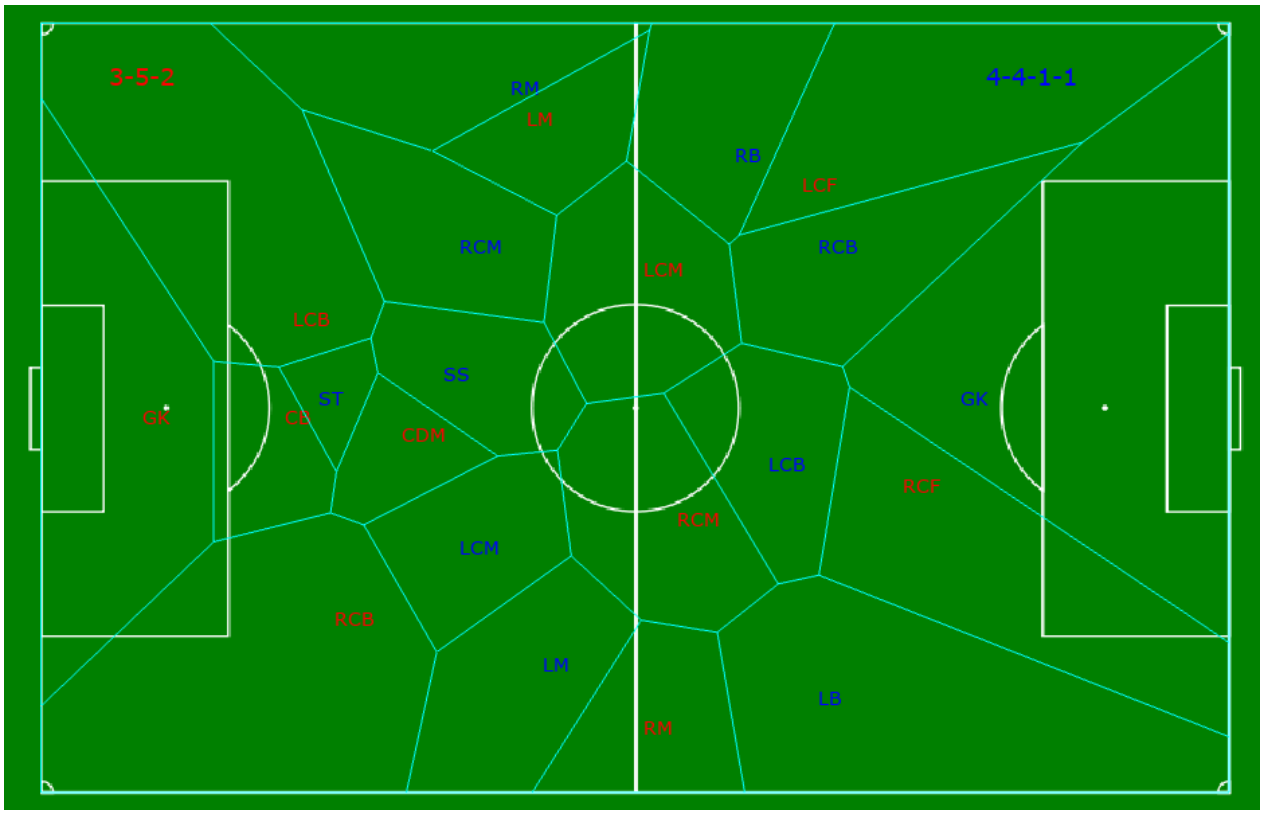

Figura 2. Diagrama de Voronoi dos Jogadores

daquele objeto do que de qualquer outro. Nos desportos, a distribuição espacial dos jogadores em campo é determinada pelo comportamento de interação estabelecido nos níveis de jogador e equipa [32] tornando os diagramas de Voronoi uma ferramenta útil para analisar partidas. Neste contexto, os diagramas de Voronoi são calculados usando a posição dos jogadores no campo (ver Figura 2). Os diagramas de Voronoi podem ajudar os treinadores a ver quão bem os jogadores usam o espaço, encontrar novos espaços para atacar e identificar áreas na defesa que a equipa deixa em aberto. Neste artigo, a área da célula de Voronoi do jogador será usada como um recurso que influencia o sucesso do passe.

2.4.1. Trabalho Relacionado

Cada vez mais no futebol moderno as equipas investem fora de campo e uma das maneiras é na obtenção de dados e na análise dos jogos, assim as ferramentas automatizadas para apoiar a análise processo são necessárias.

Num estudo realizado por Müller-Budack, Theiner, Rein et al. [33], foi introduzida uma abordagem analítica que classificava e visualiza automaticamente a formação da equipa com base nos dados de posição dos jogadores. Esta nova abordagem calculava a semelhança com base em modelos pré-definidos para diferentes formações táticas. Os resultados demonstraram que o resumo da formação visual já fornece informações valiosas e é capaz de resumir ações individuais em jogos de futebol.

Noutro estudo ([34]) foram aplicadas técnicas de aprendizagem automática para selecionar a formação tática e estratégia adequada. Foi concluído que os modelos teóricos de jogo realizados neste estudo coincidem com as tendências modernas do futebol, ou seja, além do clássico 4-4-2 o flexível e hoje popular 4-3-3 ou 4-2-3-1 rendem o melhor *payoff*. Para criar o modelo teórico realista do jogo, foi utilizado um conjunto de dados contendo a pontuação final e as formações das equipas de mais de 25.000 jogos.

2.5. Resumo

Neste capítulo podemos verificar que o gesto técnico passe é um dos mais importantes e mais utilizados no futebol e que existem várias áreas para estudar, sendo uma delas os fatores que influenciam o sucesso ou não. Sendo o passe uma ação conjunta perceber as dinâmicas da equipa, como a sinergia é útil para compreender a complexidade do gesto técnico.

Também na aprendizagem automática e IA já existem imensas aplicações nos diferentes ramos do futebol do campo tático a aspetos como a venda de bilhetes ou análise de desempenho de jogadores.

Como já foi referido existem vários trabalhos dentro da área do gesto técnico passe, a sinergia e aprendizagem automática. Este trabalho irá contribuir para perceber como ligar estes três conceitos e perceber quais os fatores que influenciam o passe, fatores esses como por exemplo a área do Voronoi ou o jogador mais próximo. Através deste capítulo, que corresponde à primeira etapa da metodologia CRISP DM, foi atingido um objetivo da dissertação que passava por compreender os conceitos fundamentais no futebol que são usados para perceber as dinâmicas das equipas (ver Tabela 1).

Tabela 1. Avaliação dos Objetivos

Objetivos	Resultados
Compreensão dos conceitos fundamentais no futebol	Cumprido
Ferramenta de IA que ajude na compreensão do sucesso ou não do passe	
Estudo e aplicação de diferentes técnicas de aprendizagem automática para modelação do sucesso do gesto técnico passe	Não Cumprido
Avaliação da aplicabilidade à compreensão do gesto técnico passe	Não Cumprido
Identificação dos fatores chave no sucesso do passe	
Visualização da relação das variáveis explicativas com a variável alvo	Não Cumprido
Determinação das variáveis mais importantes no sucesso do passe	Não Cumprido

CAPíTULO 3

Identificação dos Fatores de Sucesso no Passe (Metodologia CRISP DM)

A elaboração desta dissertação propõe, como referido anteriormente, perceber as variáveis e investigar os fatores que influenciam o sucesso do passe. Para perceber e chegar aos objetivos propostos foram feito três iterações da metodologia CRISP DM, em cada uma delas foram feitas várias experiências através de implementação de árvores de decisão e máquinas de vetor de suporte.

3.1. Influência da "Área do Voronoi" e da "Distância à baliza" (1º iteração CRISP DM)

Através de uma primeira reunião com o especialista foram definidas as variáveis e as técnicas a utilizar para chegar aos objetivos propostos. Esta primeira iteração é composta por duas experiências com os mesmos dados e variáveis, a diferença entre as duas experiências é a técnica utilizada (árvores de decisão e máquinas de vetor de suporte). Em ambas as experiências é primeiro realizada uma etapa de preparação dos dados para chegar aos dados e variáveis necessárias. De seguida, na modelação, são implementadas as duas técnicas de aprendizagem automática referidas e por fim é feita a avaliação dos resultados através da análise dos parâmetros de avaliação das técnicas e também análise subjetiva do especialista.

3.1.1. Compreensão dos Dados

Na primeira etapa, onde é dado ênfase à compreensão dos dados e como foram recolhidos os dados, foi realizada uma reunião com o especialista, um treinador de futebol, para perceber melhor os dados recolhidos e a sua dimensão. A base de dados contém 40 693 entradas e 16 variáveis. Os dados recolhidos são todos de jogos referentes a partidas da Primeira Liga Francesa. Cada registo corresponde a uma ação técnica praticada durante o jogo (exemplo passes ou remates). As variáveis presentes são as seguintes:

- *Jogo* (Inteiro): Identificador único do jogo;
- *Período*(1,2): Se é a primeira parte (1) ou segunda parte (2) da partida;
- *Tempo* (decimal) : tempo (em segundos) ;
- Nome: Identificação único do jogador que executa o evento;
- *Equipa* (EQP , ADV) : Equipa que tem a posse de bola;
- *Direção do Ataque* (-1,1): direção do ataque da equipa que tem a posse de bola;
- *Zona da Bola* (I , E): se a zona da bola está numa área exterior (E) ou interior (I) do Voronoi (ver Figura 3) ;
- *Zona da Bola Centroide* (1, 2 , 3 e 4): zona onde está a bola em relação ao centróide da equipa e aos quatro quadrantes, podemos ver na Figura 3 onde se situa cada número;
- *Resultado* (EMPATA, GANHA, PERDE): resultado atual da equipa que tem a posse de bola;
- *Evento*(caracteres): ação técnica praticada;
- *Área Total* (decimal, m^2): área total do Voronoi do jogador que realizou a ação;
- X (decimal): posição longitudinal do jogador, sendo a coordenada (0,0) o centro do campo. Neste campo o máximo é 52 e o mínimo -52;
- Y (decimal): posição lateral do jogador, máximo 36 e mínimo -36
- *Distância à baliza*(decimal): distância do jogador que tem a bola até à baliza adversária;
- *Continuação*: a bola continua na posse de bola da equipa (0) ou passa para o adversário (1);
- *Distância à baliza* (decimal): distância entre o jogador que tem a bola e a baliza adversária, em metros.

3.1.2. Preparação dos Dados

Nesta etapa foram escolhidas em conjunto com o especialista três variáveis, para realizar as primeiras técnicas. As variáveis utilizadas foram a *Área Total*, *Distância à baliza* e *Continuação*. O objetivo desta primeira análise era saber se a continuação da posse de

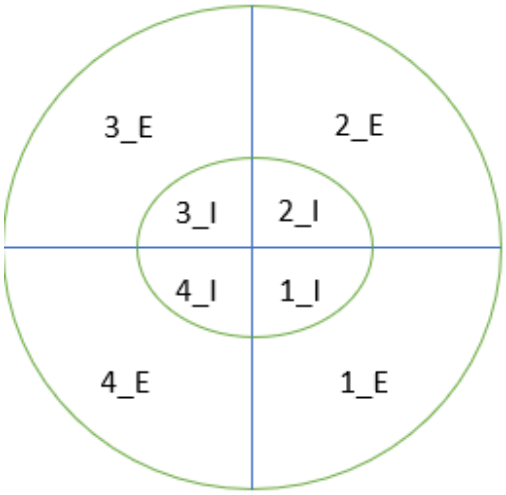

Figura 3. Divisão do centroide nos quatro quadrantes

bola na execução de um passe dependia da distância à baliza e da área total do Voronoi do jogador que executa o passe.

Para chegar ao objetivo, primeiramente foi feita uma limpeza à base de dados de modo a ter somente as variáveis de interesse. A coluna *Evento* foi filtrada pelo evento "passe". Após estas alterações foram obtidos 15 224 eventos. Desses eventos, em 84% a equipa continuava com a posse de bola (classe 0 na coluna *Continuação*) e somente em 16% a posse de bola era perdida para a outra equipa (classe 1 na coluna *Continuação*).

Nos dados que ficam para analisar pelas técnicas de automatização podemos verificar que a média da área do Voronoi é 195 m^2 e da distância à baliza adversária é 61,551 m. Destes dados de média podemos ver que, sendo o comprimento do campo de 105 m a média dos passes situa-se no campo defensivo e com uma área relativamente grande daí a maioria dos passes terem sucesso. As medianas destas duas variáveis são muito semelhantes ás médias (106,330 m^2 área do Voronoi e 60,852 m na distância à baliza adversária).

3.1.3. Modelação

Nesta etapa são selecionadas e avaliadas as técnicas a utilizar. Para esta análise foram utilizadas duas técnicas de aprendizagem automática: árvores de decisão e máquinas de vetores de suporte.

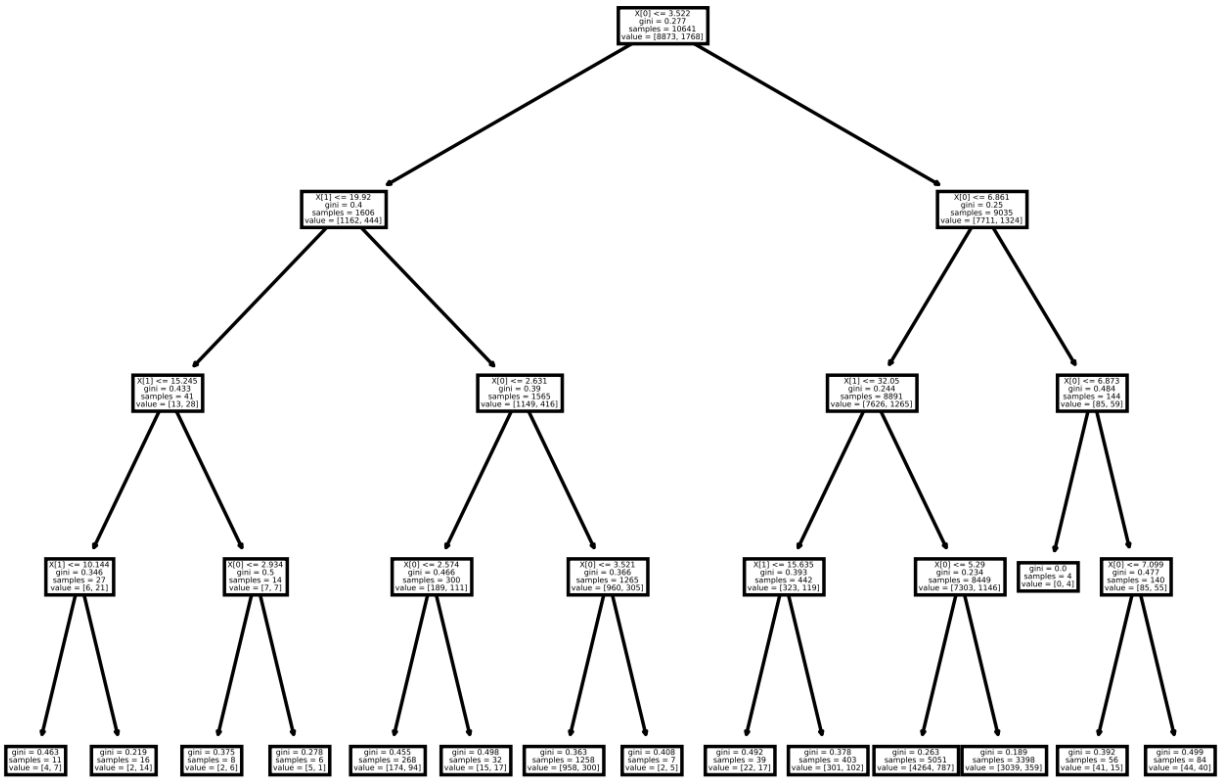

Figura 4. Árvore de Decisão - 1ª Iteração

3.1.3.1. *Árvores de Decisão* Para a técnica da árvore de decisão foi feita uma árvore com quatro níveis e o critério de seleção foi o índice de *Gini*. Este critério mede o grau de heterogeneidade dos dados,logo, pode ser utilizado para medir a impureza de um nó. Quando este índice é igual a zero, o nó é puro o que indica que só contém uma classe. Por outro lado, quando ele se aproxima do valor um, o nó é impuro (aumenta o número de classes distribuídas neste nó). Quando, nas árvores de classificação com partições binárias, se utiliza o critério de Gini tende-se a isolar num ramo os registos que representam a classe mais frequente.

No primeiro nível o critério utilizado foi área do Voronoi superior a 52,776 m^2. Neste primeiro nível a divisão foi de 76,2% para o nó em que a área era superior e somente 23,8% para área inferior. Aqui podemos ver alguma falta de equilíbrio nos dados. Na Figura 4 podemos analisar a árvore de decisão. Através da árvore foi construido o gráfico presente na Figura 5 para ver as divisões dos níveis da árvore de decisão através dos

pontos. De referir, que na Figura 5 as medidas da área total do Voronoi correspondem ao logaritmo base 10 da medida presente na base de dados e na árvore de decisão. Tanto pelo gráfico como pelo diagrama de árvores de decisão podemos verificar que os eventos com classe 0 (a equipa não continua com a posse de bola) são poucos e todos os pontos estão concentrados num curto espaço. Através desta técnica não foi possível tirar uma conclusão sobre a relação da área do Voronoi e distância à baliza adversária com o desfecho do passe devido ao que foi dito anteriormente. Na Tabela 2 podemos verificar os valore da taxa de acerto, precisão, sensibilidade e o valor do F1.

Tabela 2. Resultados Árvores de Decisão 1º Iteração

Árvore	Taxa de Acerto (Accuracy)	Precisão (Precision)	Sensibilidade (Recall)	Valor do F1
4 níveis	0,826	0,357	0,032	0,058

Ao analisar estes valores concluímos que apesar do valor da taxa de acerto ser aceitável as outras métricas não, principalmente o valor da sensibilidade que é muito baixo o que significa que o modelo vai errar muitas vezes nos casos em que a bola não contínua na posse da equipa. O valor elevado da precisão pode ser devido à falta de equilíbrio da variável de saída.

3.1.3.2. *Máquinas de Vetor de Suporte* A próxima técnica utilizada foi Máquinas de Vetores de Suporte. Numa primeira análise o *kernel* utilizado foi o linear. Devido à distribuição dos pontos, tal como analisado anteriormente, este *kernel* não teve sucesso pois o algoritmo não encontrou nenhuma forma de dividir os pontos de forma linear. Com a analise da Figura 5 onde podemos ver os pontos de ambas a classe muito próximas e concentrados na mesma área do gráfico é fácil perceber porque razão não foi encontrada forma de dividir os pontos linearmente

Foi decidido então utilizar o *kernel* radial. Na Tabela 3 podemos ver as métricas avaliadas no modelo.

Tabela 3. Resultados Máquina de Vetores de Suporte 1º Iteração

Modelo	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
SVM	0,835	0,692	0,012	0,023

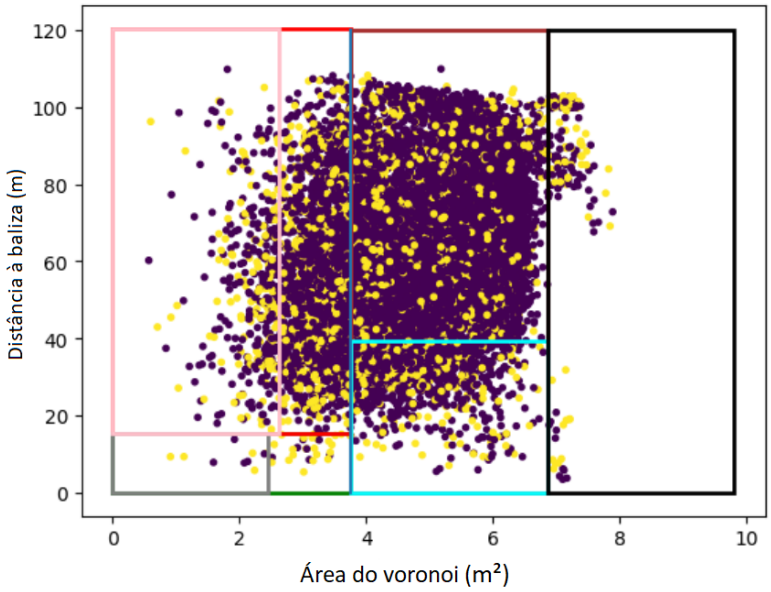

Figura 5. Gráfico com níveis de árvore de decisão (amarelo o passe não teve sucesso, roxo o passe teve sucesso)

Tal como na primeira experiência os valores da taxa de acerto são razoáveis mas os outros valores são precisamente o contrário. Estes maus resultados são muito por causa da dispersão entre as classes utilizadas. Nos gráficos a seguir podemos analisar isso. Na Figura 6 é mostrado como o *kernel* dividiu os pontos, sendo que a cor amarela é quando a equipa não continua com a posse de bola. É difícil através deste gráfico encontrar um padrão ou concluirmos em que zonas é mais propício a perder a posse de bola.

De seguida, no gráfico presente na Figura 7 foram colocados os pontos dos passes por cima do gráfico com as zonas que a máquina de suporte vetorial indicou como propício ao passe ter sucesso (zona roxa) e ao passe não ter sucesso (zona amarela). De relembrar que os valores da Área do Voronoi são referentes ao logaritmo natural dos valores da base de dados para uma melhor visualização dos pontos.

Nos dois gráficos podemos ver as duas zonas, zona amarela que significa que houve interrupção da posse de bola (1) e zona roxa que a equipa continua com a posse de bola. Ao analisarmos estes dois gráficos não é possível tirar grandes conclusões sobre qual influência da área do Voronoi e a distância à baliza adversária na interrupção da posse de bola da equipa. A única análise que podemos ver é que quando a distância à baliza é

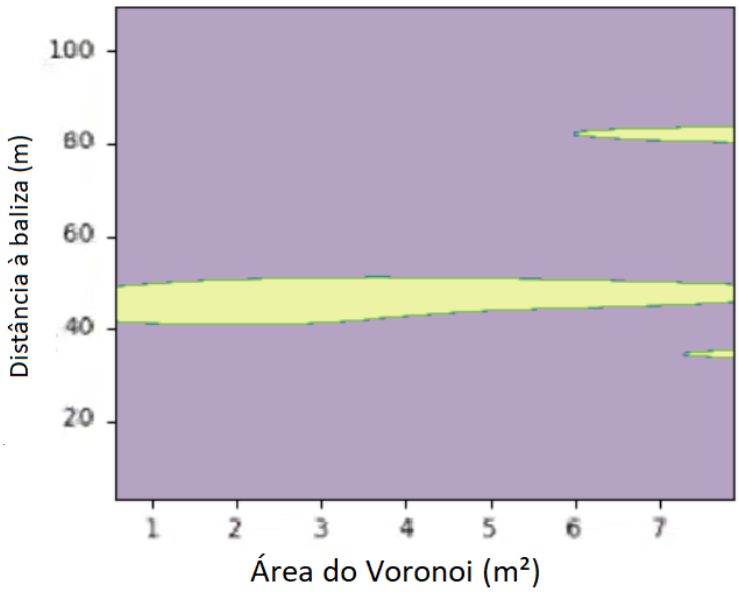

Figura 6. Gráfico da máquina suporte vetorial da 1ª iteração (zona roxa passe com sucesso, zona amarela passe sem sucesso)

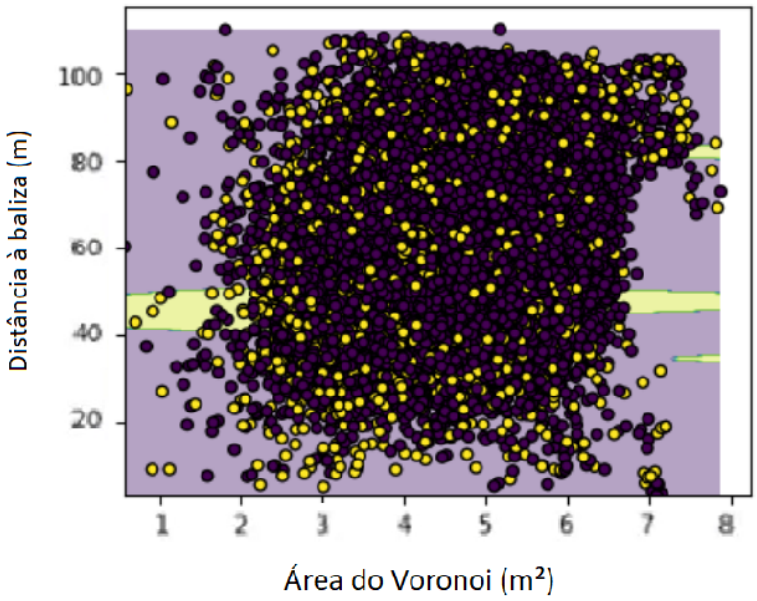

Figura 7. Gráfico da máquina suporte vetorial da 1ª iteração com os pontos referentes aos passes (zona roxa passe com sucesso, zona amarela passe sem sucesso)

menor a probabilidade de perder a bola aumenta, pois é nas zonas de menor distância à baliza que estão concentradas as zonas de interrupção da posse de bola.

3.1.4. Avaliação

Com a implementação das duas técnicas referidas e depois de feita a sua análise podemos dizer que os objetivos não foram atingidos na totalidade e que não foi possível tirar grandes conclusões sobre a dependência das variáveis e quais as variáveis mais importantes. Na Tabela 4 podemos ver o resumo dos objetivo e se foram cumpridos ou não. Em relação ao objetivo da implementação da ferramenta de IA que ajude na compreensão do sucesso do passe, os resultados das métricas de avaliação utilizados não foram suficientes para concluir que a ferramenta ajuda na tomada de decisão. Em relação ao outro objetivo, não foi possível perceber qual é realmente a influência da área do Voronoi e da distância à baliza no sucesso do passe, pois os pontos do sucesso ou não do passe estavam muito espalhados por todas as áreas do gráfico.

Após outra reunião com o especialista foram analisados os resultados. Na técnica de árvores de decisão foi dito pelo especialista que este método não indicava em nada o que era a sua intuição e a técnica era bastante questionável. Na sua opinião os resultados não encaixam na sua sensibilidade para os dados, acrescentando que faria mais sentido que a distância à baliza fosse o primeiro critério a ser usado na divisão dos dados. Quanto aos resultados obtidos através das máquinas de vetor de suporte o especialista diz que faz sentido quanto mais perto da baliza estiver o jogador maior ser a probabilidade de perder a bola. O especialista estranha é que o Voronoi não tenha relação com o sucesso do passe. Depois desta discussão como um dos problemas era a disparidade entre o número de vezes que a equipa perde a bola e que continua com a posse de bola foi discutido que se deveria utilizar somente o penúltimo e último evento da jogada. Com isto os valores de 1 e 0 iriam equilibrar.

3.2. Influência do penúltimo e último evento (2ª Iteração CRISP DM)

A segunda iteração da metodologia é composta também por duas experiências, que dizem respeito ás duas técnicas de aprendizagem automática abordadas anteriormente. As varáveis de entrada e saída serão as mesmas mas os dados serão diferentes de modo

Tabela 4. Avaliação da Primeira Iteração

Objetivos	Resultados
Compreensão dos conceitos fundamentais no futebol	Cumprido
Ferramenta de IA que ajude na compreensão do sucesso ou não do passe	
Estudo e aplicação de diferentes técnicas de aprendizagem automática para modelação do sucesso do gesto técnico passe	Cumprido
Avaliação da aplicabilidade à compreensão do gesto técnico passe	Não Cumprido
Identificação dos fatores chave no sucesso do passe	
Visualização da relação das variáveis explicativas com a variável alvo	Cumprido
Determinação das variáveis mais importantes no sucesso do passe	Não Cumprido

a equilibrar os valores da variável de saída (continuação ou não da posse de bola). Para equilibrar os valores os eventos que irão ser analisados serão o penúltimo e último de cada jogada. Tal como na iteração anterior é feita a preparação dos dados, de seguida a modelação e por fim a avaliação dos resultados das técnicas utilizadas.

3.2.1. Compreensão dos Dados

Como foi dito anteriormente, nesta iteração vão ser utilizadas as mesmas variáveis de entrada (área do Voronoi e a distância à baliza adversária) e a mesma variável de saída (interrupção ou não da posse de bola).

Como estas variáveis estão presentes na base de dados da iteração anterior, os materiais foram exatamente os mesmos. A base de dados é composta pelas 18 colunas referidas anteriormente e por 40 693 entradas.

3.2.2. Preparação dos Dados

Tal como na iteração anterior nesta etapa foi feita a limpeza à base de dados de modo a existir somente as variáveis escolhidas a utilizar nas técnicas.

Apesar da limpeza anteriormente, em que filtramos pelas colunas que iriam interessar para a análise (área do Voronoi, distância à baliza e interrupção da posse de bola) , foi preciso também adicionar uma coluna à base de dados de modo a chegar ao último e penúltimo evento. Esta coluna indicava quais eram o penúltimo ou último evento consoante se no evento a seguir se perdia ou não a posse de bola.

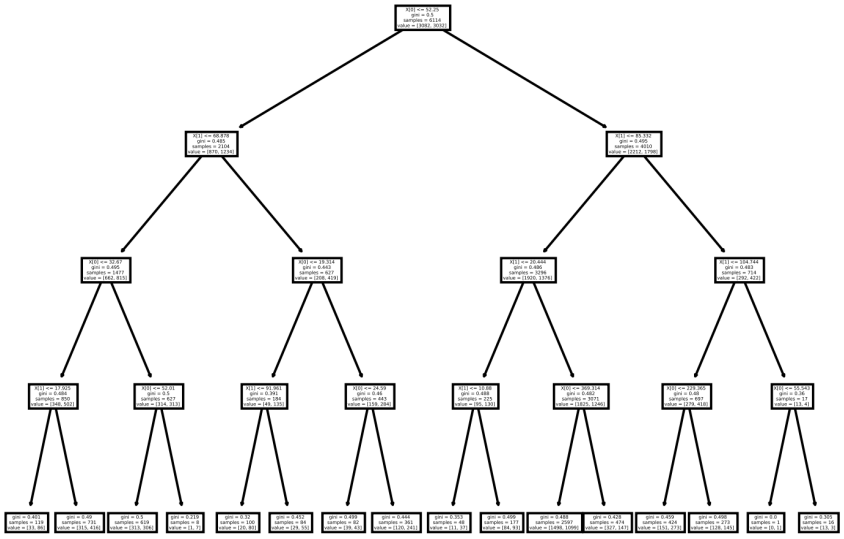

Figura 8. Árvore de Decisão - 2ª iteração

Após esta primeira limpeza e numa primeira leitura dos dados foi necessário retirar um *outlier*, a linha em questão dizia que a área total do Voronoi era 0 o que se pode concluir que era um erro do sistema. Como nesta iteração somente utilizamos o penúltimo e ultimo evento a divisão entre interrupção ou continuação da posse de bola é aproximadamente 50% para cada. O número total de eventos foi de 8 735, em que 4 368 a equipa continuou com a posse de bola.

3.2.3. Modelação

Nesta etapa são selecionadas e avaliadas as técnicas a utilizar. Para esta análise foram utilizadas duas técnicas de aprendizagem automática: Árvores de Decisão e Máquinas de Vetores de Suporte.

3.2.3.1. *Árvore de Decisão* Para a técnica da árvore de decisão foi desenhada uma Árvore com 4 níveis e o critério de seleção foi o índice de Gini.

No primeiro nível o critério foi se a área do Voronoi é superior a 52,25m^2. Neste primeiro nível a divisão foi de 65,4% para o nó em que a área era superior e somente 34,6% para área inferior. Este primeiro nível é uma área muito estreita mas divide as amostras em 65%. Podemos ver a divisão utilizada pela árvore na Figura 8. Na Tabela 5 estão os valores das métricas utilizadas para avaliar a árvore.

Tabela 5. Resultados Árvores de Decisão - 2ª Iteração

Modelo	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
Árvore (4 níveis)	0,570	0,523	0,562	0,540

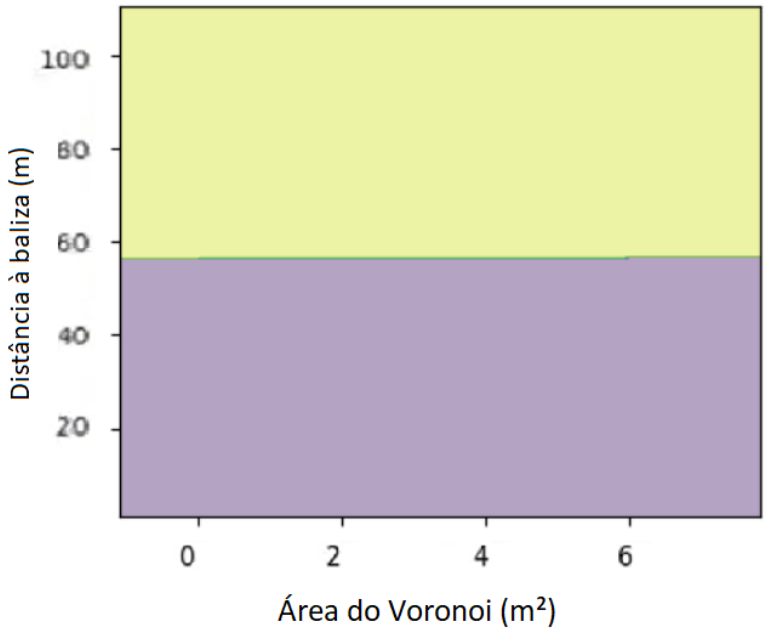

Figura 9. Gráfico máquina suporte vetorial entre as variáveis área do voronoi e distância à baliza - 2ª iteração (zona amarela os passes com sucesso e zona roxa os passes que não tiveram sucesso)

Apesar do valor da taxa de acerto ser mais baixa do que as experiências anteriores os restantes valores já têm valores mais aceitáveis. Esta melhoria das métricas claramente é devido ao equilíbrio dos valores utilizados pela classe.

3.2.3.2. *Máquinas de Vetor de Suporte* Tal como na iteração anterior a próxima técnica utilizada foi as máquinas de vetores de suporte. Como os valores de interrupção e não interrupção da posse de bola estão em percentagens iguais foi utilizado com sucesso o *kernel* linear. Nesta técnica podemos observar na Figura 9 que os pontos são divididos por um valor muito perto dos 53 m da distância à baliza. Na Tabela 6 estão os valores das métricas que foram utilizadas para avaliar a máquina de vetores de suporte.

Tabela 6. Resultados Máquinas de Vetor de Suporte - 2ª Iteração

Modelo	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
SVM	0,574	0,573	0.509	0,537

Os valores são muito próximos da experiência anterior utilizado nesta iteração. A taxa de acerto tem valores baixos mas os restantes subiram em relação à anterior máquina de vetores de suporte.

3.2.4. Avaliação

Nesta iteração após a modelação e a interpretação dos resultados podemos dizer que os objetivos voltaram a não ser atingidos, podemos ver o resumo dos objetivos na Tabela 7. Como nas Árvores de Decisão e nas Máquinas de Vetores de Suporte o nível de precisão e taxa de acerto foram baixos não se pode concluir que estas ferramentas ajudem na tomada de decisão. Também não foi possível saber que influência as variáveis utilizadas têm no sucesso ou não do passe.

Após uma reunião com o especialista, para avaliar e compreender melhor os resultados das técnicas utilizadas, foi concluído que estes resultados fazem sentido porque a diferença entre o penúltimo e último evento é quase nula pois são efetuados num espaço de tempo muito curto. Concluindo, esta nova experiência resolveu o problema da anterior que era normalizar a variável de saída mas surgiu um outro problema que é não haver diferença nos dados entre essas variáveis.

Para a próxima iteração para resolver algum destes problemas vai ser utilizado outra base de dados e mais variáveis. Para compreender melhor a sinergia coletiva, nomeadamente as ligações interpessoais faladas anteriormente, vão ser utilizadas todas as posições dos jogadores em campo e não só os dados do jogador que tem a bola.

3.3. Influência da Posição dos Jogadores (3ª iteração CRISP DM)

A terceira iteração da metodologia é composta por três experiências, nessas três experiências vão ser construídas árvores de decisão para um conjunto diferente de variáveis. Na primeira experiência vão ser utilizadas as coordenadas posicionais e área do Voronoi dos 22 jogadores em campo, totalizando assim 66 variáveis. Os jogadores irão ter um

Tabela 7. Avaliação da Segunda Iteração

Objetivos	Resultados
Compreensão dos conceitos fundamentais no futebol	Cumprido
Ferramenta de IA que ajude na compreensão do sucesso ou não do passe	
Estudo e aplicação de diferentes técnicas de aprendizagem automática para modelação do sucesso do gesto técnico passe	Cumprido
Avaliação da aplicabilidade à compreensão do gesto técnico passe	Não Cumprido
Identificação dos fatores chave no sucesso do passe	
Visualização da relação das variáveis explicativas com a variável alvo	Cumprido
Determinação das variáveis mais importantes no sucesso do passe	Não Cumprido

identificador da equipa ((f) sinaliza a equipa foco e o (o) sinaliza a equipa oponente. Na segunda experiência são somente utilizados as variáveis que dizem respeito ao jogador que executa o passe e por último na terceira são utilizadas as variáveis relativas ao jogador que executa o passe em conjunto com as do jogador mais próximo da equipa adversária e da sua equipa. Tal como na iteração anterior é feita a preparação dos dados, de seguida a modelação e por fim a avaliação dos resultados das técnicas utilizadas.

3.3.1. Compreensão dos Dados

Nesta segunda iteração a base de dados corresponde a anotações e dados posicionais de 13 jogos da Primeira Liga Francesa (Ligue 1). Esta base de dados contém 563.067 entradas e 11 variáveis. Cada entrada corresponde a uma ação técnica realizada no jogo; as variáveis correspondem ao posicionamento do jogador e outros atributos que descrevem a ação técnica (por exemplo, um passe). Estes incluem os seguintes:

Jogo: (inteiro): identificador único do jogo;

Período: (1,2): primeira parte (1) ou segunda parte (2) do jogo.

Tempo: (decimal): tempo (em segundos).

Equipa: (f, o): identificador da equipa a que pertence o jogador, equipa (f)oco ou (o)ponente.

Missão tática: (classe): missão tática do jogador (e.g., GK - guarda-redes, LB - defesa esquerdo, ST - avançado).

X: (decimal): posição no eixo longitudinal (em metros), sendo a coordenada (0,0) o centro do campo. Neste campo o máximo é 52 e o mínimo -52;

Y: (decimal) : posição no eixo lateral (em metros), máximo 36 e mínimo -36

Área do Voronoi: (decimal): área do Voronoi do jogador (em m^2).

Evento: (classe): ação técnica executada (e.g., Passe, Remate).

Distância: (decimal): distância, em metros, do jogador que tem a bola até à baliza adversária.

Zona da Bola: (E, I): se a bola está numa zona (E)xterior ou (I)nterior da área do Voronoi.

Continuação: (0, 1): a bola continua na posse da equipa (0) ou muda para o adversário (1).

Ângulo: (decimal): Ângulo, em radianos, à baliza adversária.

Além das informações referentes a cada evento outras informações de anotação também foram utilizadas, nomeadamente a formação tática predominante adotada por cada equipa (na Figura 2, um 3-5-2 para a equipa foco vermelho, e 4-4-1-1 para a equipa adversária, azul).

Usando as variáveis (x,y e área do voronoi), quatro experiências foram realizadas. Em todas as experiências foram considerados todos os passes (10 332) feitos no primeiro tempo de cada um dos 13 jogos. Destes, apenas 1 405 (13,6%) não foram bem-sucedidos (esses dados muito pouco equilibrados representam um desafio para as técnicas de aprendizagem automática).

3.3.2. Preparação dos Dados

Tal como na iteração anterior nesta etapa foi feita a limpeza à base de dados de modo a existirem somente as variáveis escolhidas a utilizar nas técnicas. Como esta iteração é composta por três experiências, em cada uma das experiências foi feita uma preparação e limpeza de dados de modo a obter as variáveis desejadas. Na primeira experiência são utilizadas três variáveis (x, y e a área do Voronoi). Na segunda experiência são utilizadas variáveis do mesmo tipo mas apenas as que correspondem ao jogador que executa o passe. Por fim, na última experiência são utilizadas as variáveis do jogador que executa o passe e dos jogadores mais próximos da sua equipa e da equipa adversária. Para obter os jogadores mais próximos foi utilizada a distância euclidiana. Em todas

estas experiências são utilizados os 10 332 passes executados nos 13 jogos analisados. Dos passes executados têm sucesso 1 405, ou seja, 13,6%.

Em todas as experiências a preparação dos dados é feita somente através de uma filtragem de colunas.

Na segunda experiência é feita uma filtragem do jogador que tem a bola através da coluna *Zona da Bola*, o jogador que executa o passe é o único que tem esta coluna preenchida que diz respeito à zona do Voronoi (E,I) onde se encontra a bola. Depois de feita esta filtragem é feita a filtragem das colunas que interessam analisar.

Na terceira experiência são utilizados os dados da segunda experiência ao qual se juntam os jogadores mais próximos. Para encontrar o jogador mais próximo foi utilizado a distância euclidiana entre os jogadores, onde guardamos os valores dos jogadores da equipa adversária e da própria equipa que tivessem menor valor.

3.3.2.1. *Comparar Jogos* Os passes ocorrem dentro de um contexto: um jogo. Como a dissertação envolve diferentes jogos, é importante avaliar quantitativamente quão (di)semelhantes dois jogos são. A similaridade de cosseno é um método bem conhecido que pode ser usado para isso. Para avaliar quão semelhantes duas partidas são, cada partida é descrita por um vetor de atributos (digamos $\mathbf{a}$ e $\mathbf{b}$) e o valor de semelhança $\mathrm{sim}_{\cos}(\mathbf{a}, \mathbf{b})$ calculado através da Equação 3.1.

$$\mathrm{sim}_{\cos}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{3.1}$$

Na dissertação, três tipos diferentes de atributos de vetores são usados para caracterizar cada jogo:

- Tática da equipa: um jogo é descrito pela formação adotada por cada equipa. As formações da Figura 2 são descritas pelo vetor

$$\underbrace{[3, 5, 2, 0}_{Foco}, \underbrace{4, 4, 1, 1]}_{Adversrio}.$$

- Missão tática: um jogo é descrito pela missão tática (por exemplo, GK, ST) dos jogadores em campo

$$\underbrace{[GK, LB, LCB, \ldots, ST, SS}_{Focos}, \underbrace{GK, LB, LCB, \ldots, ST, SS]}_{Adversrio}$$

O valor 1 é usado se um jogador com essa missão tática estiver em campo, 0 caso contrário. O exemplo da Figura 2 é descrita pelo vetor

$$\underbrace{[1,0,0,\ldots,0,0}_{Foco},\underbrace{1,1,0,\ldots,1,1]}_{Adversrio}$$

- Importância das Variáveis: uma correspondência é descrita pelo valor de importância, σ_i calculado para o recurso i da árvore de decisão.

$$\underbrace{[\sigma_1,\sigma_2,\sigma_3,\ldots,\sigma_N]}_{Nvariveis}$$

3.3.3. Modelação

A modelação vai ser composta por três experiências, cada uma com diferentes variáveis. Primeira com todos os jogadores em campo, a segunda somente com o jogador que executa o passe e por último o jogador que efetua o passe mais o jogador mais próximo da sua equipa e da equipa adversária. Também nesta etapa são analisadas quatro métricas: Taxa de Acerto, Precisão, Sensibilidade e Valor do F1.

Experiência com todas as variáveis

A fim de identificar as variáveis mais relevantes para o sucesso do passe, a primeira experiência foi realizada usando as variáveis (**x**, **y** e **Área Voronoi**) de todos os 22 jogadores em campo (66 variáveis no total) e como valor de saída o resultado do passe. Foram criadas duas árvores de decisão, uma com 6 níveis (representada na Figura 10) e outra com 20 níveis.

Tabela 8 apresenta o valor da avaliação para a árvore de decisão.

Tabela 8. Resultados Árvores de Decisão - 3ª Iteração

Árvore	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
6 níveis	0,820	0,158	0,083	0,108
20 níveis	0,790	0,235	0,500	0,161

Analisando os resultados podemos ver que embora a árvore de decisão com 6 níveis tenha uma taxa de acerto maior as outras métricas são muito baixas. Usando o método da importância das variáveis, foi calculada a importância das treze variáveis utilizadas. As variáveis foram ordenadas de acordo com sua importância, utilizando valores máximos e

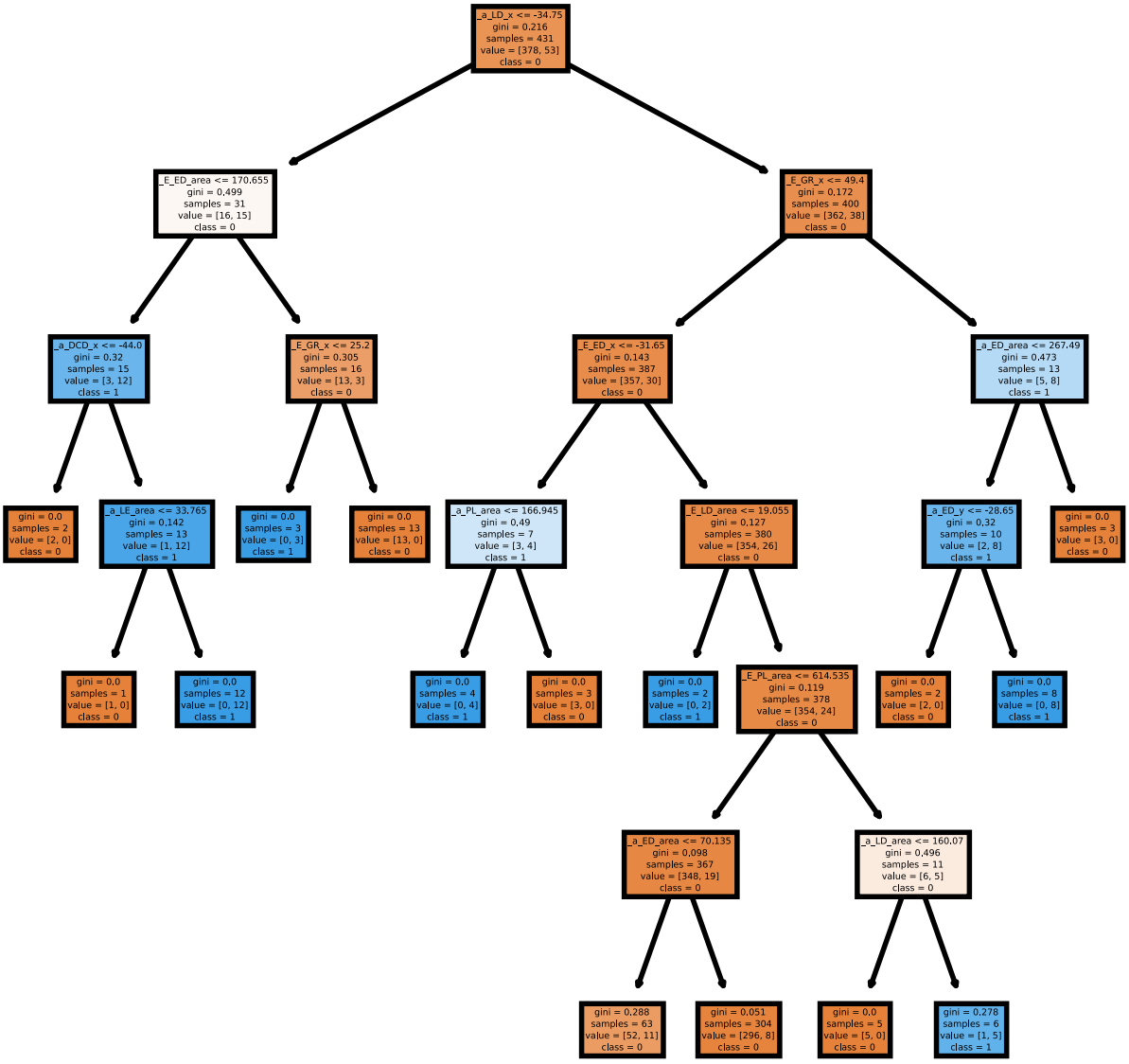

Figura 10. Árvore de decisão com variáveis de todos os jogadores em campo

médios nas partidas em que estavam presentes. A Tabela 9 apresenta as cinco principais variáveis em ambos os critérios. A Figura 11 mostra a importância das 66 variáveis em cada partida, ordenadas pela média. Das 66 variáveis nem todas aparecem em todos os jogos e algumas presentes num determinado jogo podem não ser utilizadas pela árvore de decisão. Como são muitas variáveis e não é possível colocar de forma legível o nome das mesmas no gráfico, foi construída a Tabela 9 de modo a perceber quais as que têm mais influência por valor máximo numa árvore e por valor médio nas treze árvores .

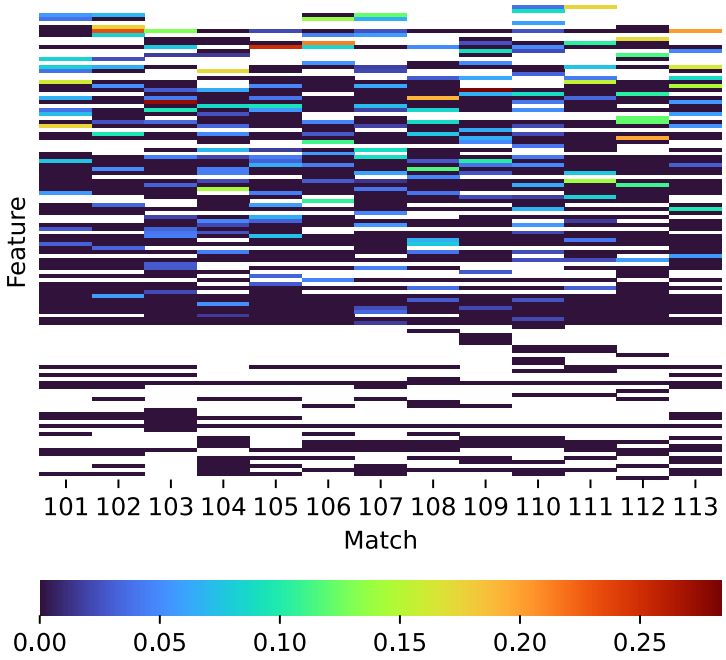

Figura 11. Importância das variáveis nos jogos (por média), azul menor relevância, vermelho maior relevância, branco variável não presente devido ao facto dessa posição tática não ser utilizada nesse jogo

Tabela 9. Top 5 variáveis mais influentes nas árvores de decisão por valor máximo e por média e número de jogos em que se encontra presente

Principais variáveis por valor máximo		Principais variáveis por média		
Variáveis	Valor	Variáveis	Valor	Jogos
f_GK_x	0,29	o_CM2_area	0,11	2
o_RCB_x	0,27	f_RM_x	0,09	1
f_RB_x	0,24	f_CAM_x	0,06	4
o_LCB_x	0,23	f_CAM_y	0,06	4
o_CAM_y	0,21	f_LM_y	0,06	1

Analisando o ranking por média, nos primeiros lugares aparecem características que não aparecem em muitos jogos (entre um a quatro jogos). Os valores apresentados são baixos, o que pode indicar que não há uma variável comum que se destaque em todas as partidas. Nomeadamente, nenhuma das diferentes classes de variáveis, **x**, **y** e **Área Voronoi** pode ser considerada dominante. Por outro lado, todas as cinco principais variáveis estão associadas a missões táticas do meio campo (quatro pertencem à equipa de foco).

Considerando o valor máximo para a importância das variáveis nenhuma delas tem um valor muito alto. Isso indica que a importância das variáveis em todos os jogos é muito dispersa. Analisando as equipas podemos ver que duas são relativas à equipa foco e três são relativas à equipa adversária, nenhuma das equipas tem destaque na influência. No entanto, a classe de variáveis **x** tem destaque, bem como as missões táticas defensivas. Esta dispersão foi analisada por computação analisando a semelhança entre a importância das variáveis para todos os pares de correspondência usando a similaridade do cosseno. A Figura 12 indica uma baixa similaridade entre todos os pares de jogos. No entanto, é de notar que a maioria dos pares que apresentam maior semelhança (por exemplo, 106 – 108, 101 – 102 e 108 – 111) também apresentam uma missão tática muito semelhante.

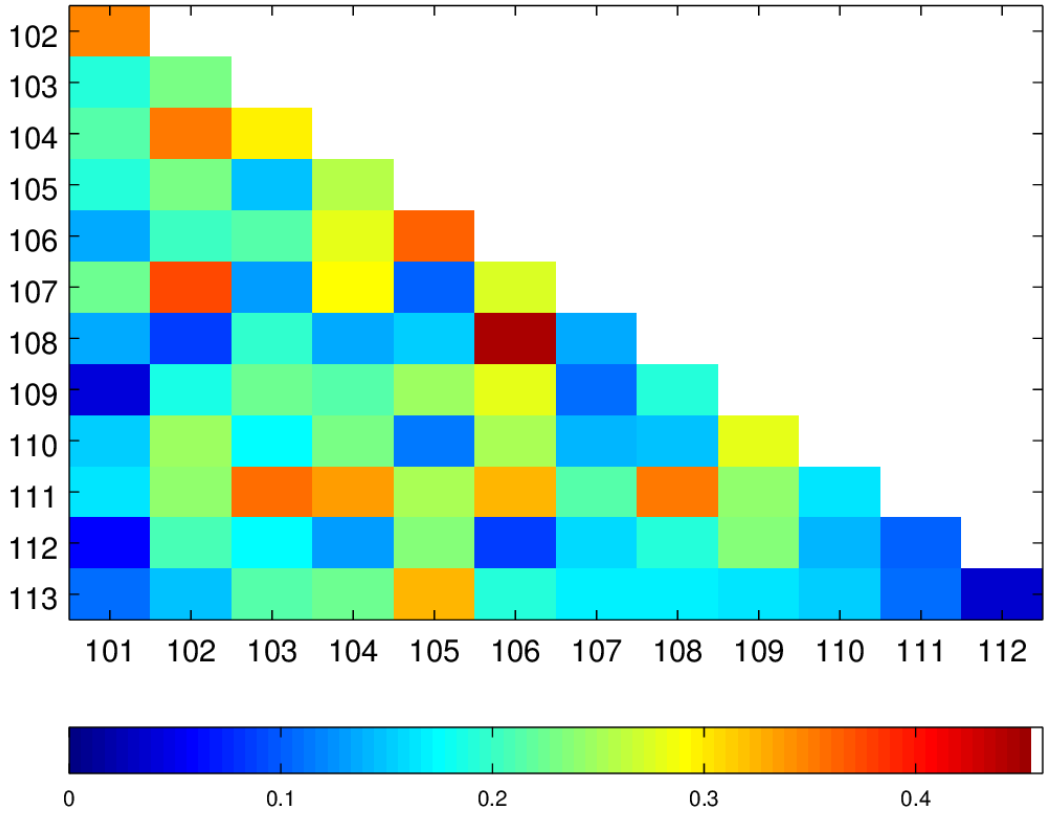

Figura 12. Semelhança da importância das variáveis entre pares de jogos usando a similiridade do cosseno

Experiência com o jogador que executa o passe

Nesta experiência foram utilizadas as seguintes variáveis do jogador que executa o passe: **x**, **y** e **Área Voronoi**, **Distância**. Como neste caso não era necessário saber a missão tática do jogador, não foi dividido por jogos. Tal como a experiência anterior foram avaliadas duas árvores com diferentes níveis (6 e 20). A árvore com seis níveis está representada na Figura 13.

Os valores avaliados para concluir sobre o desempenho das árvores estão representadas na Tabela 10. Tal como as experiências anteriores os valores da taxa de acerto foram muito bons mas os restantes são muito baixos.

Tabela 10. Resultados Árvores de Decisão do jogador com bola

Árvore	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
6 níveis	0,882	0,250	0,095	0,139
20 níveis	0,849	0,187	0,286	0,224

A variável com mais importância foi a área do Voronoi e com menos importância foi a variável y . Podemos ver todos os valores da importância das variáveis na Tabela 11

Tabela 11. Importância das variáveis do jogador com bola

Variáveis	Valor
X	0,13
Y	0,11
Área do Voronoi	0,43
Distância à Baliza	0,33

Depois da avaliação da árvore de decisão e verificado a importância de cada variável foi removida a variável com menor importância, a variável y. Ao retirarmos esta variável podemos verificar na Tabela 12 que as métricas para a avaliação existe uma ligeira melhoria mas continuam aquém do esperado, o que significa que o y não tem muita importância no processo de decisão das árvores.

Quanto à importância das variáveis, representado na Tabela 13, a área do Voronoi contínua a ser a mais importante e a variável x passou a ter mais preponderância e a distância à baliza teve um ligeiro decréscimo.

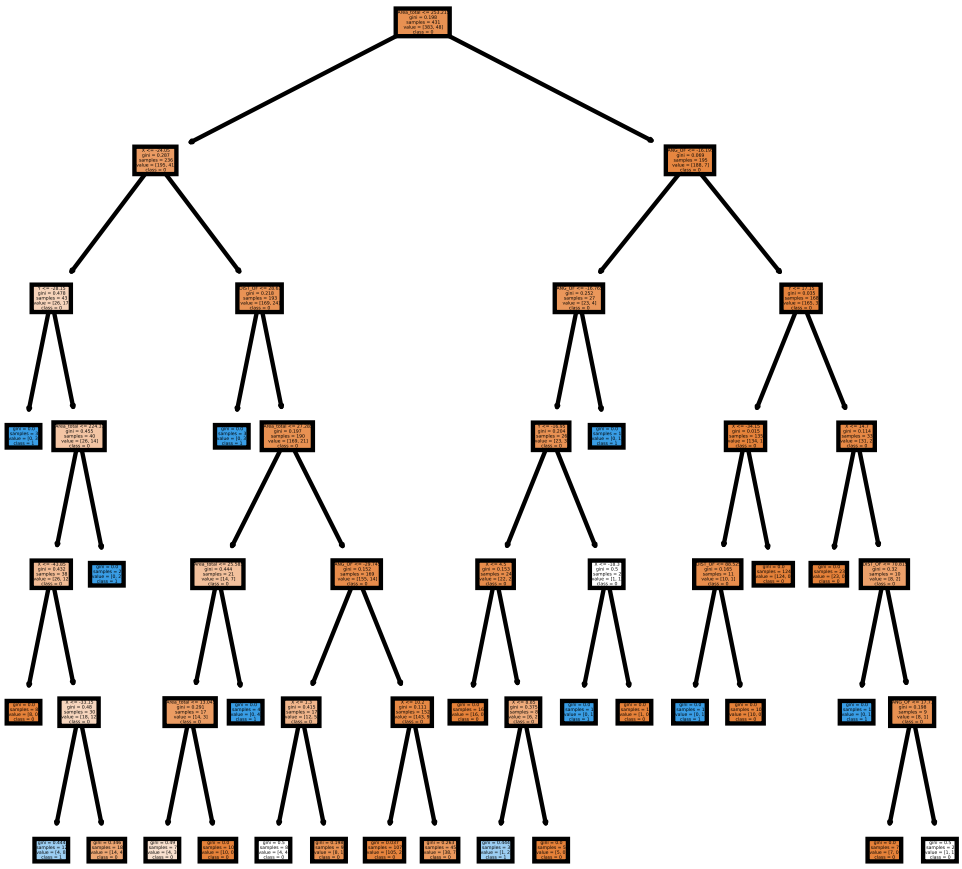

Figura 13. Árvore de Decisão do jogador com bola

Tabela 12. Resultados da Árvore de Decisão sem variável y

Árvore	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
6 níveis	0,872	0,500	0,227	0,311
20 níveis	0,849	0,187	0,286	0,224

Tabela 13. Importância das variáveis sem y

Variáveis	Valor
X	0,24
Área do Voronoi	0,54
Distância à Baliza	0,22

Experiência do jogador que executa o passe e o jogador mais próximo

Para contornar os problemas das experiências anteriores foram utilizadas as variáveis do jogador que executa o passe combinadas com variáveis do jogador mais próximo da mesma equipa e da equipa adversária. Com a introdução dos jogadores mais próximo de cada equipa ajuda a compreender a sinergia utilizada na execução do passe. Além das coordenadas longitudinais e lateral, foram consideradas as características relacionadas ao espaço "disponível"(área de Voronoi) e "suporte/pressão"(distância ao companheiro/adversário):

p_x(y): (decimal): posição longitudinal (transversal) do portador da bola.

p_area: (decimal): área do Voronoi do portador da bola.

p_dist: (decimal):distância do portador da bola à baliza.

f_sep: (decimal): distância do portador da bola ao jogador mais próximo da sua equipa.

o_sep: : distância do portador da bola ao jogador mais próximo da equipa adversária.

f(a)_area: (decimal): área do Voronoi do jogador mais próximo da equipa do portador da bola (f) e da equipa adversária (o).

f(o)_dist: (decimal): distância à baliza do jogador mais próximo da equipa do portador da bola (f) e da equipa adversária (o).

Tabela 14. Resultados Árvores de Decisão do jogador mais próximo

Árvore	Taxa de Acerto	Precisão	Sensibilidade	Valor do F1
6 níveis	0,820	0,100	0,150	0,120
20 níveis	0,780	0,310	0,320	0,032

Analisando a Tabela 14 podemos ver que embora a árvore de decisão de 6 níveis tenha maior taxa de acerto as outras métricas são menores. A Figura 14 mostra a importância de cada variável ordenada pelo valor médio crescente. A característica mais importante

é a distância entre o portador da bola e o adversário mais próximo. Segundo o especialista isto faz sentido, pois à medida que o adversário está a fazer pressão, a dificuldade do sucesso do passe aumenta. Na verdade, todas as três características relacionadas ao adversário são encontradas nos cinco primeiros lugares, reforçando a hipótese de que a interação com o adversário mais próximo é de extrema importância no sucesso do passe. Por outro lado, as características relativas ao companheiro de equipa estão entre as menos importantes, especialmente a distância até à baliza.

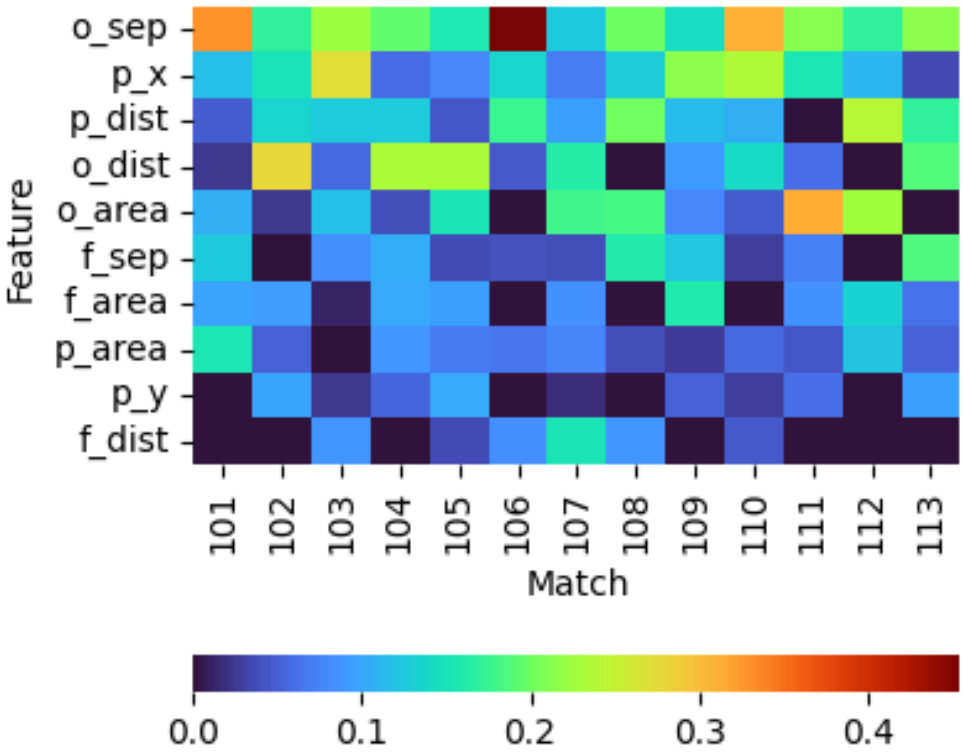

Figura 14. Importância das Variáveis do Jogador com Bola (x, área do voronoi, distância à baliza e y), Companheiro de Equipa (distância ao portador da bola, distância à baliza e área do voronoi) e Adversário (distância ao portador da bola, distância à baliza e área do voronoi) em cada Árvore de Decisão, azul menor importância, vermelho maior importância

3.4. Avaliação

Com a realização destas três experiências podemos tirar algumas conclusões em relação ás variáveis mais importantes, embora as métricas de avaliação tenham ficado, novamente, longe do pretendido tal como nas iterações anteriores. Podemos analisar a avaliação dos objetivos na Tabela 15. Em relação ao objetivo da implementação da ferramenta IA não se pode cumprir devido ao fraco desempenho com este conjunto de dados das ferramentas de aprendizagem automática utilizadas.

Com a primeira experiência verificamos que tanto a precisão como os outros fatores de avaliação ficaram longe do esperado. Ao analisarmos as variáveis com maior importância e média superior podemos tirar algumas conclusões. Primeiro que as que tem o valor mais alto pertencem a posições defensivas e exceto uma variável todas elas pertencem à coordenada x. No top 5 de média das variáveis, verificamos que todas elas tiveram presentes em poucos jogos, pertencem maioritariamente a posições do meio campo e por último, exceto uma, todas elas pertencem à equipa foco.

Na segunda experiência os valores das métricas da árvore de decisão melhoraram ligeiramente mas voltaram a ficar longe do desejado. Nesta experiência a conclusão que podemos tirar é que a área do Voronoi é muito importante na decisão tal como a distância à baliza.

Na última experiência os valores das métricas voltaram a descer ligeiramente. Em termos da importância das variáveis podemos concluir que a interação com o adversário é a mais importante para o sucesso ou não do passe, pois no top 5 está presente em 3 variáveis.

No conjunto das três experiências podemos então avaliar que ter um conjunto de variáveis que estão diretamente relacionadas com ao processo (passe) possibilita uma classificação consistente de variáveis entre os jogos. As variáveis que dizem respeito ao jogador mais próximo parecem ter uma importância muito elevada em relação a todas as outras. Ao contrário das outras iterações as áreas do Voronoi perderam importância.

Tabela 15. Avaliação da Terceira Iteração

Objetivos	Resultados
Compreensão dos conceitos fundamentais no futebol	Cumprido
Ferramenta de IA que ajude na compreensão do sucesso ou não do passe	
Estudo e aplicação de diferentes técnicas de aprendizagem automática para modelação do sucesso do gesto técnico passe	Cumprido
Avaliação da aplicabilidade à compreensão do gesto técnico passe	Não Cumprido
Identificação dos fatores chave no sucesso do passe	
Visualização da relação das variáveis explicativas com a variável alvo	Cumprido
Determinação das variáveis mais importantes no sucesso do passe	Cumprido

CAPíTULO 4

Conclusões

O principal objetivo desta dissertação era usar e compreender como diferentes técnicas de aprendizagem automática podem ajudar a avaliar as tomadas de decisão dos jogadores através de análises de dados e métricas. Através deste objetivo principal surge a questão de investigação:

- Como é que uma ferramenta de IA pode ser usada para compreender os fatores que influenciam o sucesso de um gesto técnico?

Ao longo da dissertação foram utilizadas várias métricas, bases de dados e ferramentas de aprendizagem automática para compreender o que influencia o sucesso do passe. Ao longo da dissertação foram utilizadas duas bases de dados e técnicas de aprendizagem automática (árvores de decisão e máquinas de vetor de suporte) de modo a responder aos nossos objetivos e pergunta de investigação. Um dos objetivos desta dissertação também passava por compreender o passe como uma ação conjunta e interpretar a sinergia utilizada no passe. Através desta dissertação e dos estudos analisados podemos perceber que o passe pode ser e é uma ação conjunta que não só é influenciada pelo executante mas também por todos os jogadores em campo, exemplo disso é quando o adversário mais próximo foi utilizado ganhou mais importância do que as variáveis do jogador que executa o passe.

Nas três iterações feitas pela metodologia os resultados ficaram aquém do esperado, com a avaliação das ferramentas não ser o esperado, visto que apesar da taxa de acerto em todas as experiências ser alta a sensibilidade e a precisão serem muito baixas. Apesar das limitações, nomeadamente a baixa precisão, os resultados podem ser úteis para os profissionais de futebol. Por exemplo, eles podem ajudar a projetar tarefas de treino de passe restrito (por exemplo, com distâncias representativas para oponente e companheiros de equipa).

Com esta dissertação podemos verificar que encontrar variáveis que estejam ligadas de forma muito influente ao sucesso do passe não é nada fácil, pois o sucesso ou não do passe depende de vários fatores e não de nenhum em específico e de forma tão direta. Os resultados obtidos são devido, em parte, ao desequilíbrio dos dados. As conclusões mais detalhadas desta dissertação são as seguintes:

- O desequilíbrio sucesso/insucesso do passe prejudica a avaliação do mecanismo de decisão. Na única experiência onde houve equilíbrio dos dados, a sensibilidade e a precisão aumentaram consideravelmente mas a taxa de acerto baixou.
- A importância relativa das variáveis está de alguma forma relacionada com a formação das equipas e a missão tática dos jogadores.
- O uso de mais variáveis não garante um aumento da taxa de acerto. O importante é utilizar as variáveis adequadas.
- Ter um conjunto de variáveis que são diretamente relacionadas ao processo (passe) possibilitou uma classificação consistente de variáveis entre jogos.
- A interação com o adversário mais próximo parece ser fundamental e importante para o sucesso do passe.

Apesar dos resultados não terem sido totalmente positivos foi possível tirar algumas conclusões, e como já foi referido, abrir caminhos para futuros trabalhos de investigação. Ao longo da dissertação foram superados vários desafios o que resultou nas necessárias três iterações e na mudança da base de dados. Uma contribuição significativa desta dissertação é o artigo publicado na *icSports 2022*, que já está disponível para acesso [35].

Em relação a futuros trabalhos nesta área, sugerimos:

- Explorar as técnicas para mitigar o desequilíbrio dos dados.
- Estudar sobre outras características relacionadas com a interação com os adversários.
- Investigar por que as áreas de Voronoi não são tão relevantes quanto o esperado. Uma dica é que a área completa de Voronoi pode não ser considerada como "útil".

Referências

[1] https://www.techopedia.com/definition/23371/digital-revolution, Data de Acesso: 2021-12-22, dez. de 2020.

[2] M. Ks, «Applications of Artificial Intelligence in the Game of Football: The Global Perspective,» *Researchers World - Journal of Arts Science & Commerce*, vol. 11, pp. 18–29, jul. de 2020. doi: 10.18843/rwjasc/v11i2/03.

[3] A. A. Ibrahim, «Definition Purpose and Procedure of Developmental Research: An Analytical Review,» *Asian Research Journal of Arts & Social Sciences*, vol. 1, 2016. doi: 10.9734/ARJASS/2016/30478.

[4] https://www.datascience-pm.com/crisp-dm-2/, Data de Acesso: 2021-01-10.

[5] A. Ali, «Measuring soccer skill performance: A review,» *Scandinavian journal of medicine & science in sports*, vol. 21, pp. 170–83, abr. de 2011. doi: 10.1111/j.1600-0838.2010.01256.x.

[6] P. Power, H. Ruiz, X. Wei e P. Lucey, «Not All Passes Are Created Equal: Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data,» em *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, ACM, 2017, pp. 1605–1613. doi: 10.1145/3097983.3098051.

[7] M. Kempe, D. Memmert, S. Nopp e M. Vogelbein, «Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer,» *International Journal of Sport Science*, vol. 4, pp. 35–41, 2014. doi: 10.5923/s.sports.201401.05.

[8] R. Mackenzie e C. Cushion, «Performance analysis in football: A critical review and implications for future research,» *Journal of Sports Sciences*, vol. 31, n.º 6, pp. 639–676, 2013. doi: 10.1080/02640414.2012.746720.

[9] R. Rein, D. Raabe e D. Memmert, «Which pass is better? Novel approaches to assess passing effectiveness in elite soccer,» *Human Movement Science*, vol. 55, pp. 172–181, 2017. doi: https://doi.org/10.1016/j.humov.2017.07.010.

[10] A. Tenga, L. T. Ronglan e R. Bahr, «Measuring the effectiveness of offensive match-play in professional soccer,» *European Journal of Sport Science*, vol. 10, n.º 4, pp. 269–277, 2010. doi: 10.1080/17461390903515170.

[11] A. Arbués Sangüesa, A. Martin, J. Fernandez, C. Ballester e G. Haro, «Using Player's Body-Orientation to Model Pass Feasibility in Soccer,» 2020, pp. 3875–3884. doi: 10.1109/CVPRW50498.2020.00451.

[12] B. Keskin, «The effects on soccer passing skills when warming up with two different sized soccer balls,» *Educational Research and Reviews*, vol. 10, pp. 2860–2868, 2015. doi: 10.5897/ERR2015.2444.

[13] P. Silva, D. Chung, T. Carvalho, T. Cardoso, K. Davids, D. Araujo e J. Garganta, «Practice effects on intra-team synergies in football teams,» *Human Movement Science*, vol. 46, pp. 39–51, 2016. doi: 10.1016/j.humov.2015.11.017.

[14] D. Araujo e K. Davids, «Team Synergies in Sport: Theory and Measures,» *Frontiers in Psychology*, vol. 7, set. de 2016. doi: 10.3389/fpsyg.2016.01449.

[15] J. Milho e P. Passos, «An Exploratory Approach to Capture Interpersonal Synergies between Defenders in Football,» Frontiers Media SA, 2017. doi: 10.3389/fpsyg.2016.01449.

[16] P. Esteves, D. Araujo, K. Davids, L. Vilar, B. Travassos e C. Esteves, «Interpersonal dynamics and relative positioning to scoring target of performers in 1 vs. 1 sub-phases of team sports,» *Journal of sports sciences*, vol. 30, pp. 1285–93, 2012. doi: 10.1080/02640414.2012.707327.

[17] J. Beernaerts., B. D. Baets., M. Lenoir., K. D. Mey. e N. V. de Weghe., «Analysing Team Formations in Football with the Static Qualitative Trajectory Calculus,» em *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support - Volume 1: icSPORTS*, 2018, pp. 15–22. doi: 10.5220/0006884500150022.

[18] K. P. Murphy, *Machine learning - a probabilistic perspective*, sér. Adaptive computation and machine learning series. MIT Press, 2012, isbn: 0262018020.

[19] T. M. Mitchell, «Does Machine Learning Really Work?» *AI Magazine*, vol. 18, n.º 3, p. 11, 1997. doi: 10.1609/aimag.v18i3.1303.

[20] A. J. Stimpson e M. L. Cummings, «Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms,» *IEEE Access*, vol. 2, pp. 78–87, 2014. doi: 10.1109/ACCESS.2014.2303071.

[21] Z.-H. Zhou, «A brief introduction to weakly supervised learning,» *National Science Review*, vol. 5, n.º 1, pp. 44–53, 2017. doi: 10.1093/nsr/nwx106.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay, «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. doi: 10.48550/ARXIV.1201.0490.

[23] S. Finlay, *Predictive analytics, data mining and big data: Myths, misconceptions and methods*. Springer, 2014.

[24] B. de Ville, *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. SAS Institute, 2006, isbn: 9781599943107.

[25] R. Stacey, *Towards Data Science: The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*, https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3, Data de Acesso: 2022-04-22, 2018.

[26] K. El Bouchefry e R. S. de Souza, «Chapter 12 - Learning in Big Data: Introduction to Machine Learning,» em *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, P. Škoda e F. Adam, eds., Elsevier, 2020, pp. 225–249. doi: https://doi.org/10.1016/B978-0-12-819154-5.00023-0.

[27] N. Rommers, R. Rössler, E. Verhagen, F. Vandecasteele, S. Verstockt, R. Vaeyens, M. Lenoir e E. D'Hondt, «A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players,» *Medicine & Science in Sports & Exercise*, vol. 52, p. 1, 2020. doi: 10.1249/MSS.0000000000002305.

[28] M. Jamil, A. Phatak, S. Mehta, M. Beato, D. Memmert e M. Connor, «Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football,» *Scientific Reports*, vol. 11, 2021. doi: 10.1038/s41598-021-01187-5.

[29] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi e F. Giannotti, «PlayeRank: Data-Driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach,» vol. 10, n.º 5, 2019. doi: 10.1145/3343172.

[30] S. Barra, S. M. Carta, A. Giuliani, A. Pisu, A. S. Podda e D. Riboni, «FootApp: an AI-Powered System for Football Match Annotation,» *CoRR*, vol. abs/2103.02938, 2021. arXiv: 2103.02938. URL: https://arxiv.org/abs/2103.02938.

[31] «Trends of tactical performance analysis in team sports : bridging the gap between research, training and competition,» *Revista Portuguesa de Ciências do Desporto*, vol. 9, 2009. doi: 10.5628/rpcd.09.01.81.

[32] S. Fonseca, J. Milho, B. Travassos, D. Araujo e A. Lopes, «Measuring spatial interaction behavior in team sports using superimposed Voronoi diagrams,» *International Journal of Performance Analysis in Sport*, vol. 13, pp. 179–189, 2013. doi: 10.1080/24748668.2013.11868640.

[33] E. Müller-Budack, J. Theiner, R. Rein e R. Ewerth, «"Does 4-4-2 exist?-- An Analytics Approach to Understand and Classify Football Team Formations in Single Match Situations,» 2019. doi: 10.1145/3347318.3355527.

[34] G. Dobreff, A. Pašić, B. Sonkoly e L. Toka, «The formation game in football,» em *6th Workshop on Machine Learning and Data Mining for Sports Analytics, ECML/PKDD 2019 Workshop*, 2019.

[35] H. Muacho, «The elusive features of success in soccer passes: a machine learning perspective,» *icSports 2022*, 2022.

APÊNDICE A

Iterações da Metodologia

A.1. 1ª Iteração

- *Compreensão dos Dados*: Reunião com o especialista onde foi feita uma explicação das variáveis que compõem a base de dados. Foram definidas três variáveis para realizar as primeiras análises. Na análise dos dados também e verificou que as áreas interiores normalmente tem áreas de Voronoi menores ;
- *Preparação dos Dados*: Foi feita a limpeza da base de dados para realizar as primeiras análises;
- Modelação: Executar e avaliar as duas técnicas.
- Avaliação: Reunião com o especialista onde foi feito a avaliação das técnicas e definidas variáveis para a próxima iteração.

A.2. 2ª iteração

- *Compreensão dos Dados*: A mesma base de dados que anterior;
- *Preparação dos Dados*: Foi feita a limpeza da base de dados para realizar as primeiras análises;
- Modelação: Executar e avaliar as duas técnicas;
- Avaliação Reunião com o especialista onde foi mostrado os resultados e foram avaliados. Com esta reunião o especialista definiu outra base de dados para ser utilizada na próxima iteração.

A.3. 3ª iteração

- *Compreensão dos Dados*: Reunião com o especialista para análise da base de dados. Onde a maior conclusão tirada que nesta nova base de dados, embora houvesse mais variáveis para analisar, a variável de saída continuava a não ter equilíbrio nos dados;

- *Preparação dos Dados*: Foi feita a limpeza da base de dados para realizar as primeiras análises. Nas várias experiênicas nesta iteração foram feitos vários códigos de programação para chegar a certas variáveis, por exemplo o jogador mais próximo;
- Modelação: Executar e avaliar as duas técnicas;
- Avaliação Reunião com o especialista onde foi mostrado os resultados e foram avaliados.