UNIVERSIDADE FEDERAL DE UBERLÂNDIA

FACULDADE DE ENGENHARIA ELÉTRICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Utilização do Classificador Polinomial como ferramenta de predição de resultados de partidas de futebol

RODRIGO GRASSI MARTINS

Orientador: Prof. Dr. Luciano Vieira Lima

Uberlândia

2017

Rodrigo Grassi Martins

Utilização do Classificador Polinomial como ferramenta de predição de resultados de partidas de futebol

Tese apresentada ao Programa de Pós-Graduação Stricto Sensu da Faculdade de Engenharia da Universidade Federal de Uberlândia, como recurso parcial para obtenção do título de **Doutor em Ciências**.

Área de Concentração: Processamento da Informação

Linha de Pesquisa: Inteligência Artificial

Orientador: Prof. Dr. Luciano Vieira Lima

Uberlândia

2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

M386u
2017

Martins, Rodrigo Grassi, 1983-
Utilização do classificador polinomial como ferramenta de predição de resultados de partidas de futebol / Rodrigo Grassi Martins. - 2017.
134 f. : il.

Orientador: Luciano Vieira Lima.
Tese (doutorado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Engenharia Elétrica.
Inclui bibliografia.

1. Engenharia elétrica - Teses. 2. Reconhecimento de padrões - Teses. 3. Esportes coletivos - Teses. 4. Futebol - Teses. I. Lima, Luciano Vieira, 1960-. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

CDU: 621.3

Rodrigo Grassi Martins

Utilização do Classificador Polinomial como ferramenta de predição de resultados de partidas de futebol

Tese apresentada ao Programa de Pós-Graduação Stricto Sensu da Faculdade de Engenharia da Universidade Federal de Uberlândia, como recurso parcial para obtenção do título de **Doutor em Ciências**.

Área de Concentração: Processamento da Informação

Linha de Pesquisa: Inteligência Artificial

Prof. Dr. Luciano Vieira Lima	Prof. Dr. Alexandre Cardoso
Orientador	Coordenador do curso de Pós-Graduação

Banca Examinadora

Luciano Vieira Lima, Dr. (UFU) - Orientador

Fabiano Azevedo Dorça, Dr. (UFU)

Igor Santos Peretta, Dr. (UFU)

Júnia Magalhães Rocha, Dr.ª (IFTM)

Rubens Barbosa Filho, Dr. (UEMS)

Uberlândia

20/06/2017

À minha mãe Eliana, meu irmão Ricardo, meus avós maternos José e Nair
e à memória do meu pai Maurício.
ofereço.

À minha esposa, Sabrina

e aos meus filhos Henrique e Eduardo,

dedico.

Agradecimentos

Primeiramente, devo agradecer a Deus por ter me dado a vida e oportunidade de poder conviver ao lado de pessoas maravilhosas.

Aos meus pais Eliana e Maurício *"In Memoriam"*, que sempre me apoiaram e me incentivaram em tudo que decidi realizar e que enchem minha vida de muito amor, carinho e bons exemplos.

À minha esposa Sabrina, a quem pude confiar todas as minhas preocupações, carinho, atenção, paciência, compreensão e que foi uma das principais fontes de incentivo.

Aos meus filhos Henrique e Eduardo, que me trazem muitas felicidades.

Ao meu irmão Ricardo, exemplo de pessoa, pelo apoio em todos os momentos.

Agradecimentos especiais aos meus avós maternos José e Nair e aos meus avós paternos Dorival e Wilma.

Um agradecimento especial ao Prof. Dr. Luciano Vieira Lima que tornou possível a realização deste trabalho, pela dedicação e paciência que sempre teve comigo, pela amizade e companheirismo.

Ao Prof. Dr. Alessandro Santana Martins pela amizade e pela dedicação na realização do artigo e pelas inúmeras discussões e revisões envolvendo a confecção da tese.

Aos Profs. Dr. Leandro Alves Neves e Dr. Marcelo Zanchetta do Nascimento, pela amizade, profissionalismo, dedicação e atenção na realização do artigo.

Aos amigos da IFTM que me incentivaram na realização deste trabalho.

À Universidade Federal de Uberlândia (UFU) por tornar possível a realização do meu

trabalho e aos servidores da Faculdade de Engenharia Elétrica, pelo apoio constante e informação adquiridos durante os meus estudos.

Ao meu tio Edson, minha tia Sandra e meus primos Marcel, Murilo, Márcio e Evelyn, minha cunhada Viviane e seu esposo Fernando, a bisavó Iracema, meu sogro Bosco e minha sogra Marise pelos bons momentos de convivência.

Ao São Paulo Futebol Clube por despertar o interesse no futebol, tema desta tese.

“ Viver

E não ter a vergonha

De ser feliz

Cantar e cantar e cantar

A beleza de ser

Um eterno aprendiz ”

Gonzaguinha

Resumo

O interesse de tantas pessoas no mundo pelo futebol gera não apenas telespectadores, mas também muitas movimentações financeiras em torno desse esporte. Sistemas computacionais que trabalham com a predição de resultados e auxiliem a minimizar os riscos e maximizar os lucros tornam-se então uma importante ferramenta de trabalho para o dia a dia do futebol. A hipótese de trabalho desta tese é a de que o Classificador Polinomial(CP), uma técnica amplamente utilizada como classificador de padrões possa ser usada também como um algoritmo de seleção de características. De maneira a investigar a predição - aquilo que se diz antecipadamente - dos resultados das partidas de futebol, foram escolhidos os algoritmos Naive Bayes(NB), Árvore De Decisão(AD), Multilayer Perceptron(MLP), Radial Basis Function(RBF) e Supportt Vector Machine(SVM). A escolha desses classificadores é baseada no estado da arte. Para validar a eficácia do classificador polinomial foram realizados testes com os métodos: Análise de componentes principais(PCA) e Relief. Os dados utilizados para a abordagem proposta foram os resultados dos partidas de futebol, obtidos nos seguintes campeonatos: campeonato inglês temporada 2014/2015(CI 2014/15), campeonato espanhol temporada 2014/2015(CE 2014/15) e campeonatos brasileiro temporadas de 2010(CB 2010) e 2012 (CB 2012). As técnicas de validação do método utilizadas foram: cross validation e sliding window. Os resultados obtidos de acordo com a metodologia proposta mostram que o CP obteve as melhores acurácias quando comparados com os outros cinco classificadores utilizados: NB, AD, MLP, RBF e SVM. Ainda é possível afirmar que de acordo com os resultados o CP conseguiu melhorar a acurácia dos cinco classificadores com índices superiores ao Relief e ao PCA. Também é possível afirmar de acordo com os resultados apresentados na seção 6.6, que as acurácias obtidas com essa metodologia são equivalentes ou superiores aos resultados encontrados no estado da arte, variando de 0,96 a

0,99.

Palavras-chave: Classificador polinomial, reconhecimento de padrões, predição de partidas de futebol, seleção de características.

Abstract

The interest of so many people in the world for football generates not only viewers but also many financial movements around football. Computer systems that work with predicting results and help minimize risk and maximize profits make it an important tool for the day-to-day running of football. The working hypothesis of this thesis is that the Polynomial Classifier, a technique widely used as a standard classifier, can also be used as a feature selection algorithm. In order to investigate the prediction of the results of soccer matches. The Naive Bayes, decision tree, MLP, RBF and SVM algorithms were chosen. The choice of these classifiers is based on the state of the art. To validate the efficacy of the polynomial classifier, tests were carried out using the following methods: Principal Component Analysis(PCA) and Relief. The data used for the proposed approach were the results of soccer matches, obtained in the following championships: English championship season 2014/15 (CI 2014/15), Spanish championship season 2014/2015 (CE 2014/15) and Brazilian championships seasons of 2010 (CB 2010) and 2012 (CB 2012). The validation techniques used were: cross validation and sliding window. The results obtained according to the proposed methodology show that the CP obtained the best accuracy when compared to the other five classifiers used: Naive Bayes(NB), Decision Tree(DT), Multilayer Perceptron(MLP), Radial Basis Function(RBF) e Supportt Vector Machine(SVM). It is still possible to affirm that from agreement with the results the CP was able to improve the accuracy of the five classifiers with indices higher than Relief and PCA. It is also possible to state according to the results presented in section ref ArtData the accuracy obtained with this methodology is as good or superior to the results found in the state of the art, varying from 0.96 to 0.99.

Keywords: Polynomial classifier, Recognition of patterns, prediction of football matches, selection of features.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AD	*Árvore de Decisão*
CB 2010	*Campeonato Brasileiro de 2010*
CB 2012	*Campeonato Brasileiro de 2012*
CE 2014/15	*Campeonato Espanhol temporada de 2014/15*
CI 2014/15	*Campeonato Inglês temporada de 2014/15*
CP	*Classificador Polinomial*
MLP	*Multilayer Perceptron*
NB	*Naive Bayes*
PCA	*Principal Component Analysis*
RBF	*Radial Basis Function*
SVM	*Support Vector Machine*
WEKA	*Waikato Environment for Knowledge Analysis*

Capítulo 1

Introdução

1.1 Introdução

O futebol é um esporte disputado por duas equipes de onze jogadores em um campo retangular com duas traves cada uma delas posicionada em cada uma das extremidades e uma bola. O objetivo do jogo é transpor a bola entre as extremidades utilizando basicamente toques com o pé. O vencedor da partida será o time que conseguir atingir o objetivo mais vezes no jogo [33, 58].

A simplicidade, o baixo custo e a dinâmica fizeram dessa modalidade uma verdadeira febre mundial. A última copa do mundo jogada no Brasil no ano de 2014 atingiu uma audiência televisiva de mais de 3,5 bilhões de pessoas [23]. Já a audiência da última olimpíada disputada em Londres em 2012 atingiu cerca de 4 bilhões de pessoas [47]. A diferença entre os dois eventos é que na copa do mundo temos 32 países disputando uma única modalidade, enquanto nos jogos olímpicos temos 204 países envolvidos em 26 modalidades esportivas diferentes.

O interesse de tantas pessoas no mundo pelo futebol gera não apenas telespectadores, mas também muitas movimentações financeiras em torno desse esporte. A copa do mundo de 2014 gerou de lucro para a FIFA cerca de 7,5 bilhões de dólares [23]. A movimentação financeira envolve patrocínios a clubes e seleções nacionais, venda de ingressos e produtos

licenciados. Além é claro de transações de transferências envolvendo jogadores. Esse fluxo de movimentação financeira atrai uma série de investidores, que sempre visam lucrar. Sistemas computacionais que trabalham com a predição de resultados e que auxiliam a minimizar os riscos e maximizar os lucros tornam-se então uma importante ferramenta de trabalho para o dia a dia do futebol [52].

O resultado de uma partida de futebol é o personagem central de inúmeros estudos científicos. Existe um esforço muito grande para melhorar as táticas do jogo e as características de uma determinada equipe. Na literatura, existem estudos que se concentram nas previsões de jogos de futebol [11]. A previsão em partidas de futebol é composta por resultado de uma partida (vitória, empate e derrota) que pode ser utilizada para várias finalidades, incluindo apostas. Muito esforços foram dedicados para a compreensão do futebol a partir da perspectiva dos resultados preditivos. Prever os resultados é um problema difícil devido ao grande número de fatores que devem ser levados em consideração e que nem sempre podem ser representados em valores quantitativos [32]. Por exemplo, uma equipe pode dominar completamente uma partida sob alguns aspectos, como o número de finalizações certas, o número de passes certos ou a posse de bola e não conseguir marcar um gol a mais do que a equipe adversária para vencer uma partida [6].

Na literatura é possível encontrar estudos que envolvam predição de resultados de futebol e também de outros esportes coletivos. Ulmer e Fernandez [67] estudaram as técnicas Baseline, Gaussian Naive Bayes, Hidden Markov Model, Multimodal Naive Bayes, SVM, RBF, One vs All SGD para prever resultados utilizando como parâmetros os gols marcados por cada time em 10 temporadas (da temporada 2002-03 à temporada 2011-12) do Campeonato Inglês. Hucaljuk e Rakipovic [32] pesquisaram a previsão de resultados para a UEFA Champions League também através de gols marcados com o seguintes algoritmos: Naive Bayes (NB), Redes Bayesianas, Logitboost, K-Nearest Neighbours (KNN), RF e Redes Neurais Artificiais. Com o classificador SVM, Igiri [34] estudou os dados relacionados aos resultados de partidas do Campeonato Inglês.

Com uma quantidade muito maior de características, Parinaz e Sadat [49] utilizaram dados obtidos através de *scout* e também de fisiologia para analisar o time de futebol do

Barcelona no Campeonato Espanhol. Nesse trabalho, os autores descreveram uma abordagem de redes bayesianas para previsão de resultados de partidas de futebol com o software NETICA. Apesar dos bons resultados obtidos, o modelo considera apenas uma equipe para realizar a previsão do resultado das partidas. Tax e Joustra [62] usaram os seguintes algoritmos de classificação: CHIRP, LogitBoost, DTNB, FURIA, HiperPipes, J48, Naive Bayes, Perceptron e RF. Para a seleção, eles aplicaram Relief, CfsSubsetEval, e PCA para tentar determinar dentre as 65 características levantadas quais as mais relevantes para aumentar a acurácia dos classificadores. Entretanto, eles não conseguiram determinar quais seriam estas características. O estudo levou em consideração jogos do Campeonato Holandês.

Duarte et al. [20] pesquisaram os classificadores: C5.0, JRip, RF, KNN, SVM e NB para prever as partidas do Campeonato Português com as informações obtidas através de *scout* e também dados psicológicos dos jogadores das equipes o que, no entanto, não melhorou o desempenho em termos de acurácia. Por esse motivo, é um desafio investigar informações e estratégias que facilitem a previsão dos resultados dos jogos.

Outros estudos foram desenvolvidos para prever o resultado de partidas de futebol utilizando algoritmos de aprendizagem de máquinas. Esses algoritmos são ferramentas que recebem como entrada um conjunto de características e fornece como saída a previsão do resultados (vitória, empate e derrota). Existem algoritmos que podem fornecer uma resposta mais adequada ao problema [51, 55].

A decisão da utilização de aplicar o classificador polinomial para fazer a predição de resultados de jogos de futebol é baseada na sua capacidade de aprendizagem complexa. Pois ele é capaz de atuar em padrões que podem ser linearmente inseparáveis e o sucesso obtido em outras aplicações [50]. O classificador polinomial utiliza parametrização não linear que expande de maneira não linear uma sequência de vetores de entrada para uma dimensão superior e mapeia-os para uma sequência de saída desejada. Essa expansão pode melhorar a separação das diferentes classes em um espaço vetorial. Além disso, essa estratégia apresenta as vantagens de fornecer apenas um modelo para separação ótima das classes e dessa maneira pode solucionar o problema, o que não ocorre com os modelos apresentados em outros trabalhos [10, 53].

1.2 Objetivos Deste Trabalho

O objetivo deste trabalho é explorar o método Classificador Polinomial como ferramenta de predição de resultados de partidas de futebol. Esse classificador mapeia um conjunto de amostras de treinamento e classifica as amostras previamente identificadas.

Os objetivos específicos deste trabalho são:

- mostrar como o classificador polinomial se comporta na classificação de padrões de quatro bases de dados distintas envolvendo partidas de futebol;
- comparar e demonstrar a maior eficácia do classificador polinomial com a de outros classificadores utilizados no estado da arte de predição de resultados de partidas de futebol;
- analisar o classificador polinomial como um método de otimização utilizado na seleção das melhores características para predição de resultados de partidas de futebol; e
- comparar e demonstrar a maior eficácia do classificador polinomial com a de outros seletores de características utilizados no estado da arte de predição de resultados de partidas de futebol.

1.3 Estrutura Deste Trabalho

Este trabalho consiste de sete capítulos. O primeiro capítulo mostra uma introdução do uso de algoritmos relacionados a classificação de padrões voltados à predição de resultados de partidas de futebol. Em seguida são apresentados os objetivos e a estrutura deste trabalho.

O Capítulo 2 apresenta um estudo sobre o estado da arte envolvendo predição de resultados de partidas de futebol. Tais estudos nortearam a confecção dessa tese e a publicação resultante da mesma.

O Capítulo 3 apresenta um breve estudo sobre a metodologia de coleta utilizando *scout*, método multiobjetivo para selecionar componentes. Os dados coletados dizem respeito aos fundamentos executados pelos jogadores durante uma partida de futebol. A coleta desses

fundamentos é de extrema importância, pois tais dados serão utilizados como características para identificar os padrões e dessa forma realizar a classificação do resultado de uma determinada partida.

O Capítulo 4 apresenta a hipótese deste trabalho que é a de utilizar o classificador polinomial não apenas para identificar os padrões, mas também para selecionar as melhores características e com elas melhorar a acurácia de outros classificadores existentes. Mostra também uma breve descrição dos seis classificadores que serão comparados com o Classificador Polinomial. São eles: Naive Bayes, Árvore de Decisão, MLP, RBF e SVM. E os dois seletores de características que serão comparados também com o Classificador Polinomial: Análise de Componentes Principais e Relief.

O Capítulo 5 descreve a metodologia de teste para validação do método proposto neste trabalho. Apresenta inicialmente a visão geral, em seguida as base de dados utilizadas, os vetores de características extraídas das bases, as duas técnicas de validação do modelo: *Cross Validation* e *Slinding Window* e a métrica utilizada.

O Capítulo 6 mostra os resultados obtidos nos testes realizados para a metodologia proposta. Os resultados obtidos pelos classificadores com e sem a utilização de seletores de características e uma comparação com outros estudos presentes no estado da arte.

O Capítulo 7 apresenta as conclusões e as contribuições deste trabalho, a publicação e os trabalhos futuros que poderão ser desenvolvidos a partir desta tese.

1.4 Considerações Finais Deste Capítulo

Este capítulo mostrou uma introdução do uso de algoritmos relacionados a classificação de padrões voltados à predição de resultados de partidas de futebol.

O próximo capítulo apresenta um breve estudo sobre a metodologia de coleta utilizando *scout*. Os dados coletados dizem respeito aos fundamentos executados pelos jogadores durante uma partida de futebol. A coleta desses fundamentos é de extrema importância pois tais dados serão utilizados como características para identificar os padrões e dessa forma realizar a classificação do resultado de uma determinada partida.

Capítulo 2

Estado da Arte

2.1 Introdução

Este capítulo apresenta uma visão geral sobre o estado da arte para a predição de resultados de partidas de futebol. Na literatura é possível encontrar uma série de estudos cada qual com sua própria metodologia visando o acerto antecipado de resultados de partidas esportivas, não apenas no futebol foco desta tese mas também em outras modalidades esportivas. O primeiro ponto que precisa ser analisado para quem pretende iniciar estudo nessa área é a diferença entre os termos: previsão, que será abordado na seção 2.2 e o termo predição, que será descrito na seção 2.3. O estudo aqui desenvolvido envolve apenas a predição de resultados de partidas de futebol e não a previsão. A seção de predição é divida em seis subseções que compõe os elementos centrais das metodologias encontradas no estado da arte, são eles: características, base de dados, classificadores, seletores de características, técnicas de validação de método e as métricas e resultados obtidos. Finalizando o capítulo, a seção 2.4 traz as considerações finais do mesmo.

2.2 Previsão

O termo previsão, do inglês *forecasting*, significa o que se consegue antever, o que se faz antecipadamente [11]. A previsão em partidas de futebol é composta por resultado

de uma partida (vitória, empate e derrota) que pode ser utilizada para várias finalidades, incluindo apostas. Para se realizar a previsão é necessário que se crie um mecanismo através do qual se consiga antever uma situação futura [32]. Um mecanismo que se pode utilizar para realizar a previsão é projeção de fatos futuros, levando essa conta uma dimensão de tempo [62]. Essa projeção pode ser realizada através de métodos estatísticos [5, 35] ou algoritmos computacionais como em [59, 60]. Os resultados obtidos através dessa projeção são então comparados aos dados reais, ou seja com o que realmente ocorreu naquele espaço de tempo.

O objetivo de quando se trabalha com previsão é buscar uma curva que seja o mais próxima possível da curva real [5, 32]. A projeção dos dados será realizada em um conjunto de características que são utilizadas para tentar identificar quais são os fatos mais relevantes para determinar o resultado de uma partida. No ano de 2007, Bittner et al.[5] propôs um modelo matemático autoajustável para distribuição de gols em uma partida de futebol. Por meio da análise de dados envolvendo os campeonatos nacionais da Alemanha e da Copa do Mundo, mostrou-se que a distribuição dos gols não segue nenhum dos três padrões estastistícos utilizados: aproximação de Bernoulli, *generalized extreme value* e distribuição de Poisson. O modelo apresentado aprimorou o método de Bernoulli utilizando um mecanismo de auto afirmação, que utiliza o ajuste das probabilidades de um novo gol acontecer durante uma partida. Bittner et al [35], em 2008, apresenta uma nova abordagem que aprimora o mecanismo de autoafirmação em que diferenças na qualidade da aproximação dependendo das circunstâncias politicas e culturais dos campeonatos foram analisados. Para validar a hipótese utilizaram dados dos campeonatos alemães masculinos e femininos disputados no período da guerra fria.

No contexto computacional, Silva et al. [60] propuseram um modelo não markoviano para previsão de gols e classificação de campeonatos de futebol em sistema de pontos corridos. Um algoritmo genético foi desenvolvido para elaborar aproximações e prever resultados mais relevantes em relação ao modelo Guassiano. O algoritmo proposto foi avaliado sobre um conjunto de dados do campeonatos nacionais da Itália, Espanha e Brasil. Silva e Dahmen em [59] utilizaram um novo modelo computacional baseado nas informações relacionadas à capacidade de um time vencer um jogo e o histórico de seus resultados anteriores para

avaliar dados do campeonato Espanhol, Inglês, Francês e Brasileiro. Considerando que a distribuição de gols é um fator importante em partidas de futebol, mas não é a única característica relevante para decisão de uma partida, como o jogo é uma disputa coletiva e inclui uma série de conceitos táticos envolvendo uma grande variedade de fundamentos - ações que um jogador pode executar de acordo como os eventos vão ocorrendo na partida. Estudos também têm sido propostos utilizando a probabilidade de uma vitória acontecer baseada na qualidade da troca de passes de uma equipe [24].

A linha de pesquisa envolvendo previsão de resultados de partidas de futebol é bastante promissora e com muitos problemas em aberto [24]. Entretanto o foco de trabalho desta tese envolve predição de resultados de partidas de futebol. O significado do termo predição e seu respectivo estado da arte serão apresentados na próxima seção.

2.3 Predição

O termo predição, do inglês *prediction*, significa o que se consegue predizer, o que se diz antecipadamente [11]. A predição em partidas de futebol é composta por resultado de uma partida (vitória, empate e derrota) que pode ser utilizada para várias finalidades, incluindo identificar as características mais relevantes para a determinação do resultado. Para se realizar a predição é necessário que se crie um mecanismo através do qual se consiga predizer uma situação futura [11]. Não existe uma única maneira de se realizar a construção desse mecanismo entretanto existe um conjunto de elementos que irão nortear essa concepção [11, 62]. O primeiro desses elementos são as características, descritos em detalhes na subseção 2.3.1, em seguida aparece a base de dados 2.3.2, procedido pelos classificadores 2.3.3, os seletores de características 2.3.1, as técnicas de validação do método 2.3.5 e finalmente as métricas utilizadas e os resultados obtidos aparecem na subseção 2.3.6.

2.3.1 Características

Em uma partida de futebol o número de anotações que se pode fazer para que seja possível sua correta descrição é bastante elevado [11]. Se for levado em consideração outros

fatores extracampo temos novamente um alto número de componentes envolvidos. Dessa maneira a definição de quantas e quais características serão utilizadas para se realizar a predição do resultados de partidas se torna bastante complexa [62]. Na literatura é possível encontrar trabalhos que levam em consideração uma série de fatores tais como dados de fundamentos obtidos por *scout* como em [62, 66, 67], outros como em [48] são utilizados fatores de logísticas tais como distância percorrida entre duas partidas subsequentes [62] ou fatores psicológicos como em [20].

O processo de *scout* apesar de bastante difundido e utilizado não é padronizado no que diz respeito ao número de fundamentos que serão coletados [66]. Em Owramipur [48] é utilizado um conjunto de 7 fundamentos obtidos por *scout* e um conjunto de 6 características de logísticas para se obter uma correta predição. Já em Ulmer [67] são utilizadas 9 fundamentos obtidos por *scout*. Em Igiri [34] são utilizadas como características 23 fundamentos coletados por *scout*. Em Hucaljuk [32] são utilizados 10 fundamentos obtidos por *scout*. Tax [62] utiliza 19 características obtidas por *scout*, 12 de logística e 12 baseadas em sites de apostas. Enquanto que Duarte [20] utiliza um conjunto de 12 características obtidas por *scout* e 12 características envolvendo aspectos psicológicos das equipes.

É importante ressaltar mais uma vez que não existe uma padronização para as características envolvidas com os trabalhos existentes na literatura. Pode-se afirmar entretanto que as características estão ligadas ao tipo de informação que se necessita coletar sobre um ou mais aspectos que envolvem uma partida de futebol [62]. Não é possível ainda afirmar que uma outra metodologia é superior a outra mas que são complementares. Para o caso específico desta tese utilizou-se dados de fundamentos obtidos por *scout*. O conceito de *scout* e os fundamentos que foram utilizados serão apresentados no capítulo 3.

2.3.2 Base de Dados

A exemplo do que ocorre com as características não existe uma padronização quanto às bases de dados utilizadas no estado da arte [66]. Isso ocorre devido ao fato de que os dados foram financiados por clubes ou projetos de pesquisa que raramente tornam suas bases

de dados públicas [62]. Um dos projetos que disponibiliza publicamente sua base pode ser encontrado em http://www.football-data.co.uk/ e foram objetos de estudo em [62, 67]. E também farão parte da metodologia desta tese descrito em mais detalhes no capítulo 5.

O estado da arte é bastante expansivo no que diz respeito à diversidade das base de dados utilizadas. Os dois tipos clássicos de campeonato foram estudados: copas (disputadas em jogos eliminatórios) e campeonatos de pontos corridos [67]. Kou-Yuan Huang [31] explorou dados relativos à copa do mundo, por sua vez Timmaraju [63] utilizou dados relativos à copa da Inglaterra, ambas disputadas no formato de copa - competição disputada em jogos eliminatórios.

A lista de campeonatos de pontos corridos analisados é bastante ampla: De Paola [16] analisou dados do campeonato italiano, Balduck at al [1] estudou dados do campeonato belga, Tufekci [66] o campeonato turco, Heuer [30] dados do campeonato alemão, Ulmer [67] o campeonato inglês, Duarte [20] o campeonato português, Tax [62] o campeonato holandês e o espanhol. Existem ainda trabalhos que levam em conta as temporadas completas que mesclam jogos de campeonatos de pontos corridos com competições no formato de copa, conforme é possível observar em [32, 49, 48].

2.3.3 Classificadores

O número de algoritmos de classificação encontrados na literatura é bastante amplo. É possível encontrar trabalhos que utilizaram as seguintes técnicas de inteligência artificial: *Baseline, Gaussian Naive Bayes, Hidden Markov Model, Multimodal Naive Bayes, RBF SVM, Random Forest, Linear SVM, One vs ALL SGD, Baysesian Net, Log Boost, K-NN, Artificial Neural Networks* , o software proprietário NETICA, *CHIRP, DTNB, FURIA, HyperPipes, J48.* Tax [62] e Duarte[20] apresentam uma sumarização de quantos e quais classificadores aparecem no estado da arte.

Apesar das muitas abordagens, alguns classificadores aparecem com melhores resultados na literatura, conforme destacado em [20, 62]. Esses algoritmos são: Naive Bayes, árvore de decisão, MLP, RBF e SVM, sendo a base de vários trabalhos relevantes [20, 34, 62, 67].

2.3.4 Seletores de Características

Os seletores de características é um artifício pouco explorado na literatura, ou seja, poucos trabalhos fizeram uso de algoritmos de seleção de características. Tufekci [66] utilizou Relief e Wrappers para reduzir dados relativos ao campeonato turco. Duarte [20] utilizou PCA para reduzir dados do campeonato português. Enquanto que Tax [62] utilizou PCA e Relief para reduzir características dos campeonatos inglês e espanhol.

Além de ser um artifício pouco utilizado, as abordagens citadas apresentaram melhoria no que diz respeito às acurácias obtidas por esses estudos. Tufekci [66] obteve como melhores resultados as acurácias de 67,10%, 67,43% e 69,97% com o conjunto completo das características e passou para 67,68%, 69,48% e 70,87% com a redução de características utilizando Relief. Duarte [20] obteve como melhores resultados 48,6% 50,2% e 50,8% passou para 52,2%, 59,4% e 56,4% com o método de redução de características PCA. Tax [62] obteve acurácias de 51,29% e 51,94% nos melhores casos com o conjunto inicial de características e após a redução com o PCA obteve 54,48% e 54,70%.

2.3.5 Técnicas de Validação do Método

As técnicas de validação do método também são bastante diversificadas e não seguem um determinado padrão. Alguns autores utilizam *hold-out*, amostras aleatórias, *cross-validation*, *leave-one-out*, *bootstrap*, *growing window* e *slinding window*. Apesar da diversidade de propostas, o fato do interesse das predições ser proveniente de dados temporais, ou seja, só ser de real interesse à predição de fatos futuros existe uma tendência a utilização das técnicas *growing window* e *slinding window* [62].

Para dados relativos a campeonatos disputados no sistema de copa existe uma tendência à utilização da técnica *growing window*. Isso ocorre porque a quantidade de jogos restantes vai diminuindo a medida que o campeonato vai avançando. Podemos observar a aplicação dessa técnica em [31, 63]. Apesar da técnica utilizada ser a mesma cada um dos estudos traz variação na organização das janelas.

No caso de dados relativos a campeonatos disputados no sistema de pontos corridos,

a técnica mais utilizada é a *slinding window*. Apesar da tendência a aplicação do método é bastante variável sobretudo no que diz respeito ao tamanho das janelas. Alguns trabalhos optam por utilizar duas instâncias do mesmo campeonato dividindo em quatro períodos de turno sempre utilizando se os turnos iniciais para treinamento e os sequentes para testes como ocorre em [66, 30]. Em outros casos como em [20] uma única instância do campeonato é dividido em 4 partes iguais. Também utilizam essa técnica [1, 16].

As outras técnicas utilizdas *hold-out*, amostras aleatórias, *cross-validation*, *leave-one-out* e *bootstrap* não são utilizadas de forma isolada em quaisquer trabalho. Sempre aparecem na companhia de um ou duas técnicas que levam em conta o componente temporal. A utilização dessas técnicas servem muitas vezes como complemento à validação dos resultados obtidos pelas técnicas temporais [20]. Dentro desse cenário se destaca a técnica *cross-validation* utilizada em [16, 20, 62, 63, 66]

2.3.6 Métricas utilizadas, resultados obtidos e possíveis problemas em aberto

Na literatura existem duas métricas bastante utilizadas: a primeira utiliza acurácia, conforme podemos observar em [20, 62, 63, 66], a outra abordagem se dá pela pontuação e classificações obtidas por cada equipe nos modelos propostos em comparação com a pontuação e classificações reais obtidas no campeonato conforme podemos observar em [30, 31].

No entanto, parte desses estudos mostram que técnicas de inteligência artificial vêm sendo investigadas e aprimoradas na etapa de predição de resultados em diversos campeonatos de futebol. No entanto, parte desse estudos apresentam limitações como, por exemplo, o estudo [48] que é um modelo testado para um único time, o trabalho [62] que apresentou vários algoritmos de seleção de característica mas não conseguiu selecionar quais as características mais importante para a definição do trabalho, Duarte et al.[20] que utiliza como característica dados psicológicos dos jogadores sendo o mesmo um critério muito subjetivo. Ressalta-se que os valores das métricas de acurácia são baixas quando são considerados um grupo de times com valores variando de 0.52 até 0.68.

Outros estudos foram desenvolvidos para predizer o resultado de partidas de futebol utilizando algoritmos de aprendizagem de máquinas. Esses algoritmos são ferramentas que recebem como entrada um conjunto de características e fornece como saída a previsão dos resultados (vitória, empate e derrota). Existem algoritmos que podem fornecer uma resposta mais adequada ao problema [51, 55]. A decisão da utilização de aplicar o classificador polinomial para fazer a predição de resultados de jogos de futebol é baseada na sua capacidade de aprendizagem complexa. Pois ele é capaz de atuar em padrões que podem ser linearmente inseparáveis e o sucesso obtido em outras aplicações [50]. O classificador polinomial utiliza parametrização não linear que expande de maneira não linear uma sequência de vetores de entrada para uma dimensão superior e mapeia-os para uma sequência de saída desejada. Essa expansão pode melhorar a separação das diferentes classes em um espaço vetorial. Além disso, essa estratégia apresenta a vantagem de fornecer apenas um modelo para separação ótima das classes, e dessa maneira pode solucionar o problema, o que não ocorre com os modelos apresentados em outros trabalhos [10, 53]. É importante ressaltar ainda que o modelo proposto independe de equipes, campeonatos e as características são obtidas através de *scout*, com metodologia de coleta bastante definida.

2.4 Considerações Finais deste capítulo

Este capítulo apresentou uma visão geral do estado da arte. O próximo capítulo 3 irá apresentar metodologia de coleta de dados por *scout*.

Capítulo 3

Processo de obtenção de dados utilizando a metodologia de *scout*

3.1 Introdução

O processo de *scout* em esportes é amplamente utilizado para a observação, registro e análise de desempenho técnico e tático de times em partidas disputadas. Cunha [13] define o *scout* como sendo um método numérico, o qual processa dados sobre determinada equipe durante as partidas de um jogo. Em várias modalidades esportivas é objeto de estudo: basquete, vôlei, handebol, futebol americano, beisebol e futebol [3, 4, 27, 43, 46, 61].

Os dados coletados são utilizados para conhecimento da própria equipe e também para estudos de estratégias das equipes adversárias. Segundo [18], essas informações permitem elaborar um mapeamento técnico e tático da equipe nas partidas. Esse mapa possibilita mensurar quais são as principais características de uma equipe identificando as principais jogadas, os principais jogadores e organização tática de uma equipe, no entanto, não existe uma padronização sobre quais dados devam ser coletados. Alguns trabalhos tais como [41] afirmam que a quantidade de dados gerados em uma partida é muito maior do que a capacidade que um observador tem de coletar, por isso, é muito importante que se defina o que deve ser armazenado para análise de comportamento de uma equipe para partidas futuras.

A metodologia de *scout* utilizada nesse trabalho é pautada nos fundamentos que compõem o jogo. Os 27 fundamentos que fazem parte de uma partida de futebol serão detalhados na seção a seguir.

3.2 Fundamentos

O jogo de futebol efetivamente ocorre no local onde se encontra a bola. As ações executadas sobre a mesma recebe o nome de fundamento. Calimam e Ferreira [9] definiram uma série de fundamentos que podem ser utilizados para o procedimento de *scout*. Apesar da metodologia proposta ser amplamente utilizada a mesma não é consenso dentro da comunidade esportiva, pois elenca um número muito alto de fundamentos. Vendite et al. [70] também trouxeram à tona uma proposta com um conjunto de fundamentos para estabelecerem o processo de *scout*. A nova metodologia reduziu o número de fundamentos, gerando uma generalização de alguns fundamentos. Entretanto a redução proposta também não atingiu o consenso. A proposta passou a ser amplamente utilizada, principalmente em jornais e revistas especializadas. Vendite e Arruda [71] aplicaram a metodologia proposta em uma série de jogos obtendo muitos resultados interessantes que culminaram com a publicação de um livro que é uma das grandes referências para *scout* em futebol. A metodologia adotada que será detalhada a seguir é o resultado do somatório dos fundamentos presentes nas duas propostas citadas, de maneira a considerar o maior número de especificidades do jogo.

Passe Certo e Passe Errado: A situação de passe se dá quando um jogador 1 do time A procura chutar a bola para um jogador 2 do time A. Caso a bola chegue ao jogador 2 o passe é considerado certo, conforme ilustra a Figura 3.1a e 3.1b. Entretanto se a bola não chegar ao destino desejado o mesmo é considerado errado, conforme ilustra a Figura 3.1c e 3.1d.

(a) Passe Certo - Momento a

(b) Passe Certo - Momento b

(c) Passe Errado - Momento a

(d) Passe Errado - Momento b

Figura 3.1: Exemplos de execução dos fundamentos Passe Certo e Passe Errado.

Lançamento Certo e Lançamento Errado: O lançamento é um tipo específico de

passe. É um passe cuja distância percorrida pelo chute é superior a 10 metros de distância. Normalmente no lançamento a bola também costuma atingir certa altura. A qualificação entre certo ou errado também é válida, seguindo os mesmos critérios descritos para o passe, conforme podemos observar na Figura 3.2.

(a) Lançamento Certo - Momento a

(b) Lançamento Certo - Momento b

Figura 3.2: Exemplos de execução do fundamento Lançamento Certo.

Cruzamento Certo e Cruzamento Errado: O cruzamento é um tipo específico de passe, direcionado à área do goleiro adversário. A qualificação entre certo ou errado também é válida, seguindo os mesmos critérios descritos para o passe, conforme podemos observar na Figura 3.3

(a) Cruzamento Certo - Momento a

(b) Cruzamento Certo - Momento b

(c) Cruzamento Errado - Momento a

(d) Cruzamento Errado - Momento b

Figura 3.3: Exemplos de execução dos fundamentos Cruzamento Certo e Cruzamento Errado.

Assistência: A assistência ocorre em virtude de um passe certo, lançamento certo ou

cruzamento certo que resulta em uma finalização que se encontra com a posse de bola.

Bola Recebida: A bola recebida ocorre em virtude de um passe certo, lançamento certo ou cruzamento certo.

Desarme Completo e Incompleto: O desarme ocorre quando um jogador do time B tenta tomar a posse da bola que se encontra com um jogador do time A. Caso o jogador do time B obtenha êxito e fique com a posse de bola o desarme é considerado completo. Caso o jogador do time B obtenha êxito mas não fique com a posse de bola o desarme é desarme incompleto.

Bola Recuperada: A bola recuperada ocorre quando um jogador recupera a posse da bola para a sua equipe.

Drible Certo e Errado: O drible ocorre quando um jogador do time A procura vencer a marcação do jogador adversário. Caso a ação seja bem sucedida o jogador do time A fica com a posse de bola e avança sobre o campo adversário. Caso a ação não seja bem sucedida e o jogador do time perder a posse de bola o drible é classificado como errado.

Bola Perdida: A bola perdida ocorre em duas situações distintas: a primeira situação ocorre quando o jogador sofre um desarme completo de um jogador do time do jogador adversário; a segunda situação ocorre quando o jogador erra um domínio de bola no momento em que vai receber um passe certo, situação na qual o jogador tenta receber a bola, mas acaba perdendo o domínio da mesma.

Escanteio cedido e conquistado: O escanteio ocorre quando o jogador da equipe A em situação de defesa coloca a bola para fora da linha de fundo do gramado. Nesse caso a equipe A cedeu um escanteio para a equipe B que conquistou o mesmo.

Impedimento: O impedimento ocorre quando no momento de receber um passe um jogador da equipe A está à frente de todos os jogadores da equipe B à exceção de apenas um jogador. Se tal situação ocorrer uma infração é marcada e a posse da bola fica para a equipe B.

Falta Cometida e Recebida: A situação de falta ocorre quando um jogador infringe uma das regras do jogo. A falta ocorre em umas das seguintes situações: colisões entre jogadores, ofensas entre jogadores, encostar a mão na bola (exceto se o jogador for o goleiro),

pé alto e jogada perigosa. Tais situações geram dois fundamentos: a falta cometida para o jogador que cometeu a infração e falta recebida para o jogador que a recebeu.

Cartão amarelo e vermelho: O cartão amarelo e o vermelho são utilizados para advertir o jogador. O cartão amarelo serve de alerta para o mesmo, sendo considerada uma primeira advertência mais grave. Já o cartão vermelho exclui o jogador daquela partida, ficando sua equipe obrigada a terminar o jogo com um jogador a menos.

Finalização Certa, Finalização Errada e Total de Finalizações: A finalização é a conclusão de uma jogada tramada por uma das equipes. A finalização poderá resultar ou não em um gol, que é o principal objetivo do jogo. Uma finalização pode ser realizada através de um toque na bola, exceto pela utilização da mão, pode atingir ou não a meta, incluindo a baliza. Caso a mesma atinja essa área qualificamos a finalização como sendo certa. Do contrário a mesma será considerada como errada. E existe ainda situações em que ocorreu uma tentativa de finalização, como por exemplo em uma cobrança de falta onde a bola é interceptada pela barreira. O total de finalizações inclui as finalizações certas e erradas e as tentativas de finalização.

Gol: O gol é o resultado de uma finalização certa. Nessa situação a bola transpôs a meta, gerando o gol e a alteração do resultado do jogo.

Gol Contra: O gol contra ocorre quando um jogador faz um gol para a equipe adversária alterando assim o resultado do jogo.

Defesa: O goleiro é um jogador diferenciado dentro de uma partida de futebol. A regra permite que ele possa utilizar as mãos para realizar uma jogada dentro da sua área. Além disso, a participação de um goleiro em uma partida de futebol tende a ser passiva, pois ele só entra em ação se for exigido pela equipe adversária. O principal papel que o goleiro desempenha em um jogo é o de impedir que a equipe adversária faça um gol. O atributo específico do goleiro é a defesa. O goleiro executa uma defesa quando ele impede que a bola chutada à sua meta seja convertida em gol para a equipe adversária.

3.3 Metodologia de Obtenção de Dados

A obtenção de dados segue uma metodologia padrão definida por profissionais da área de educação física. A coleta de dados de um determinado jogo é realizada por duas pessoas, chamadas de scoutistas. Cada scoutista é responsável pela obtenção de dados de um dos dois times envolvidos na partida, os dados coletados envolvem o fundamento realizado. A Figura 3 apresenta uma jogada completa, com uma sequência de fundamentos que serão coletados.

(a) Jogada - Momento A

(b) Jogada - Momento B

(c) Jogada - Momento C

(d) Jogada - Momento D

Figura 3.4: Exemplo de uma jogada com uma sequência de fundamentos.

A Figura 3.4a ilustra o primeiro momento da jogada. Nesse momento o Jogador Robinho, efetua um desarme completo. A jogada transcorre e o mesmo jogador, Robinho, efetua um drible certo, conforme podemos observar na Figura 3.4b. Ainda com a posse de bola, Robinho efetua um passe certo(3.4c), dado que a jogada resultou em um gol o passe certo

efetuado pelo jogador Robinho irá gerar uma assistência. Como resultado do passe certo, a bola chega então a um novo jogador nesse caso, Luís Fabiano, o mesmo então realiza uma finalização certa 3.4d. A finalização certa resulta em um gol gerando então o quinto momento da jogada que corresponde a uma nova entrada de dados. A sequência descrita até o momento é o resultado de ações efetuadas pela equipe A, no caso o Brasil obtidas pelo scoutista 1. A mesma jogada também irá gerar dados executados pela equipe B e capturados pelo scoutista 2. A Figura 3.4a ilustra o jogador Miguel, efetuando uma bola perdida.

A tabela 3.1 a seguir apresenta a sequência completa de dados coletadas pelos dois scoutitas, nela é possível perceber que num intervalo de tempo de 12 segundos foram executados 7 fundamentos durante toda a execução da jogada. Visto que o tempo total de jogo é de 90 minutos, ou seja 5400 segundos é possível perceber que o volume de dados para uma partida inteira será bem denso.

Tabela 3.1: Sequência de fundamentos que compõem uma jogada.

Jogador	Fundamento
Robinho	Desarme Completo
Robinho	Drible Certo
Robinho	Passe Certo
Robinho	Assistência
Luis Fabiano	Finalização Certa
Luis Fabiano	Gol
Miguel	Bola Perdida

3.4 Considerações Finais Deste Capítulo

Este capítulo apresentou a metodologia de coleta de dados por *scout*. Esses fundamentos estão presentes nas 4 bases de dados e seus respectivos vetores de características que compõem a metodologia de testes desta tese e serão apresentadas no capítulo 5.

Capítulo 4

Classificadores e Seletores de Características

4.1 Introdução

Este capítulo apresenta a hipótese de trabalho desta tese que é a de que o Classificador Polinomial, uma técnica amplamente utilizada como classificador de padrões possa ser usada também como um algoritmo de seleção de características. A seção 4.2 apresenta os detalhes desta proposta. A Seção 4.3 apresenta um conjunto de classificadores utilizados para a predição de resultados de partidas de futebol. Um conjunto de algoritmos de seleção de características também amplamente utilizados no estado da arte é apresentado na Seção 4.4. E finalmente a Seção 4.5 apresenta as considerações finais deste capítulo.

4.2 Classificador Polinomial

Esta seção apresenta o Classificador Polinomial(CP) e toda a fundamentação teórica do método será apresentada em detalhes na subseção 4.2.1 enquanto os detalhes da implementação utilizada serão descritos na subseção 4.2.2.

4.2.1 Fundamentação Teórica

O CP é um método de classificação supervisionado que apresenta resultados relevantes na análise de dados de imagens [17], principalmente em problemas cujos dados não são linearmente separáveis.

Este algoritmo procura expandir o espaço dos vetores de entrada em uma dimensão espacial maior, de forma a permitir uma separação mais adequada entre as classes analisadas. Portanto, define-se a função polinomial, conforme equação 4.1

$$y(\mathbf{x}) = \mathbf{a}^T p_n(\mathbf{x}) \tag{4.1}$$

onde $\mathbf{a}$ é o vetor dos coeficientes da base polinomial, $p_n(\mathbf{x})$ é a função da base polinomial e n a ordem ou o grau da função polinomial.

Por exemplo, dado um vetor de entrada bidimensional $\mathbf{x} = [x_1 x_2]^T$, os elementos de $p_2(\mathbf{x})$ resultam em parâmetros semelhantes aos mostrados em Campbell et al [10].

$$p_2(\mathbf{x}) = [1 \quad x_1 \quad x_2 \quad {x_1}^2 \quad x_1 x_2 \quad {x_2}^2]^T. \tag{4.2}$$

Na primeira etapa, deve-se transformar o conjunto de características d-dimensional em um vetor de base polinomial L-dimensional. Os coeficientes do polinômio são calculadas utilizando o método do mínimos quadrados. Portanto, a saída dada por $y_i = y(\mathbf{x})$ é obtida após uma combinação linear dos termos expandidos $p_n(\mathbf{x})$ através do vetor de coeficientes do polinômio definido pela equação 4.1. Em seguida, um vetor de teste deve ser expandido em termos da base polinomial, conforme exemplo mostrado na equação 4.2. Este processo é realizado por meio do produto escalar a partir do vetor expandido com o vetor de coeficientes polinomiais obtido a partir do conjunto de características d-dimensional na fase de treinamento. O sinal algébrico de $y(\mathbf{x})$ permite determinar qual classe pertence ao vetor analisado.

A regra de decisão para apenas duas classes ω_1 e ω_2, é definida pela Equação 4.3

$$\text{Decide} \begin{cases} \omega_1, & \text{se} \quad y(\mathbf{x}) > 0 \\ \omega_2, & \text{se} \quad y(\mathbf{x}) < 0 \end{cases}. \tag{4.3}$$

A matriz $\mathbf{X}$ representa o conjunto de dados de entrada, onde N é o número de padrões utilizados na fase de construção dos vetores de base polinomial L-dimensional.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}. \tag{4.4}$$

A expansão da base polinomial para cada amostra, matriz M de dimensão NXL, é determinada por 4.5:

$$\mathbf{M} = \begin{bmatrix} p_n(\mathbf{x}_1)^T \\ p_n(\mathbf{x}_2)^T \\ \vdots \\ p_n(\mathbf{x}_N)^T \end{bmatrix}. \tag{4.5}$$

A solução para as equações simultâneas é simplificada com uma notação, dada pela Equação 4.6:

$$\mathbf{Ma} = \mathbf{b}. \tag{4.6}$$

onde $\mathbf{b}$ é um vetor para o qual os elementos são todas constantes aleatórias de valor 1 ou -1, de acordo com a classe padrão de entrada. Quando existem mais equações que incógnitas, não existe uma única solução, assim sendo, deve-se procurar um vetor que minimize o erro entre $\mathbf{Ma}$ e $\mathbf{b}$. Este vetor é dado pela equação 4.7:

$$\mathbf{e} = Ma - \mathbf{b}. \tag{4.7}$$

Como o problema de minimização de erros é um problema clássico, conhecido na literatura como mínimos quadrados, o problema para $\mathbf{Ma} = \mathbf{b}$ é tratado pela equação 4.8

$$\mathbf{M^T Ma} = \mathbf{M^T b}. \tag{4.8}$$

A vantagem da Equação (4.8) é que a matriz $\mathbf{M}^T\mathbf{M}$ é uma matriz quadrada de dimensão $L \times L$ e geralmente não singular, assim pode-se encontrar uma única solução $\mathbf{a}^*$ conforme a Equação 4.9

$$\mathbf{a}^* = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{b} = \mathbf{M}^\dagger\mathbf{b}. \tag{4.9}$$

onde a matriz $\mathbf{M}^{\dagger}$ de dimensão $L \times N$ como mostrado na Equação (4.10)

$$\mathbf{M}^{\dagger} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \tag{4.10}$$

é chamada de *pseudoinversa* de $\mathbf{M}$ [21].

Nesta tese, o conjunto de características com 3 dimensões foi usado para bases polinomiais de 4ª ordem. Neste procedimento, combinações diferentes das características foram concatenadas nas 3 dimensões para obter o resultado final. A Figura 4.1 apresenta um exemplo de separação com características analisadas pelo algoritmo polinomial.

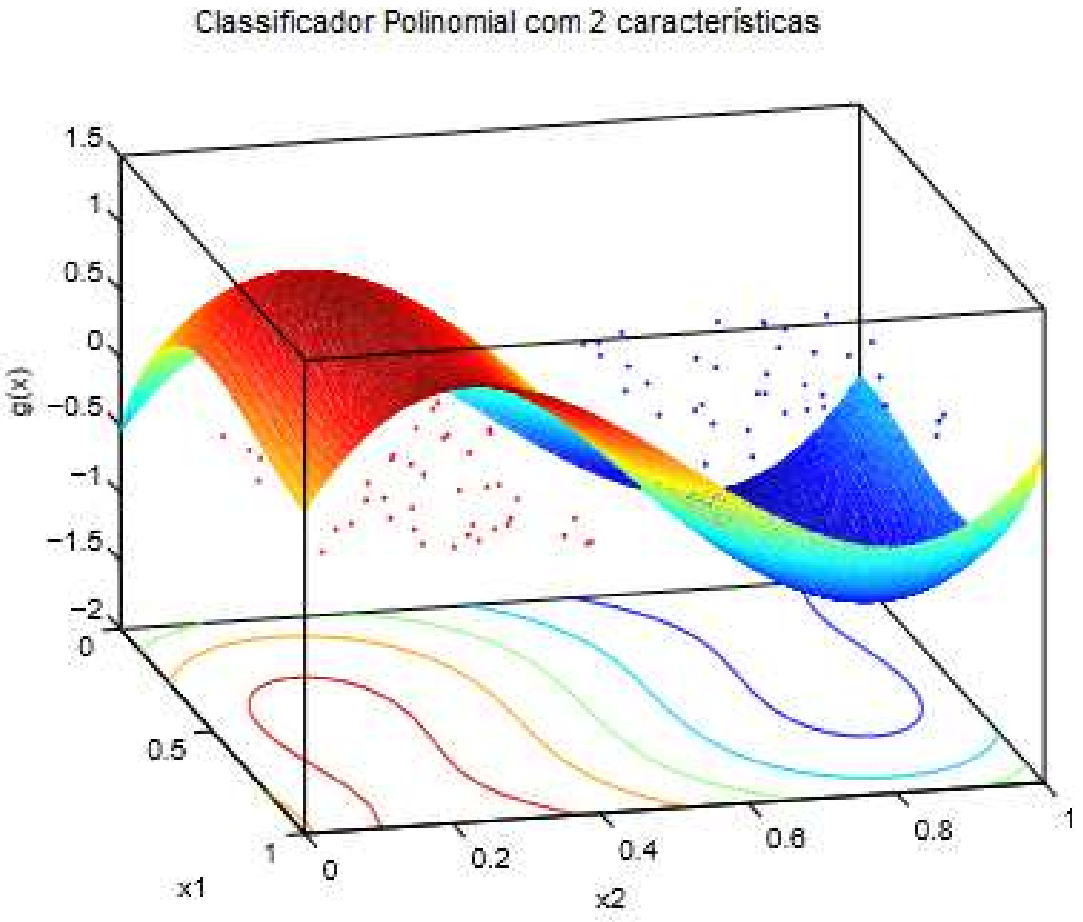


Figura 4.1: Classificador polinomial com 2 características.

A hipótese desta tese é de que o classificador pode atuar não só como um algoritmo de classificação mas também como um algoritmo de seleção de características. Para o caso

deste trabalho utiliza-se o conjunto de características com 3 dimensões para melhorar a performance de outros classificadores, a Figura 4.2 mostra o fluxograma de funcionamento do classificador polinomial em suas fases de treinamento e teste.

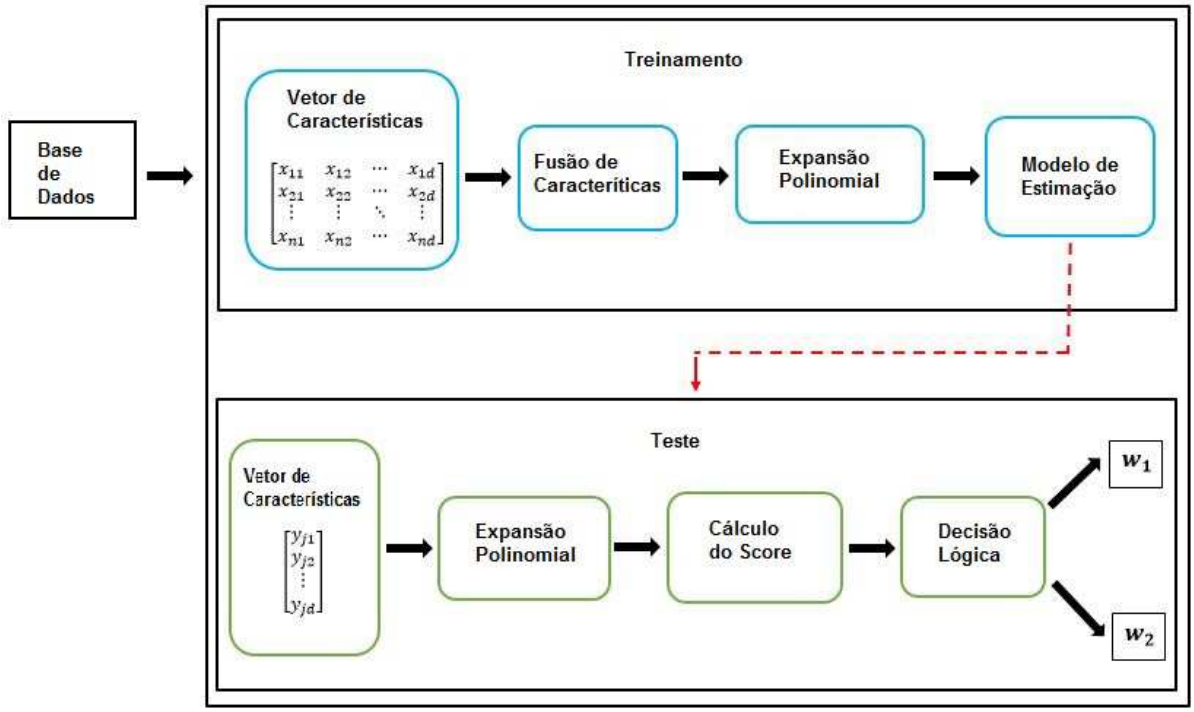


Figura 4.2: Fluxograma do funcionamento do classificador polinomial.

4.2.2 Detalhes da Implementação

.

O CP foi implementado no software MATLAB R2012b. A figura 4.3 apresenta um pseudo-código da implementação da mesma. O algoritmo possui como entrada de dados: o vetor de características x, cujo os elementos serão apresentados na seção 5.4, o grau representado pela variável n, para o caso desta tese $n = 4$ e o vetor b com os rótulos, para o caso deste trabalho $b = 1$ para a classe $(w1)$ e $b = -1$ para a classe $(w2)$. A saída do CP é a classificação em w_1 e w_2, para o caso deste trabalho vitória, empate ou derrota, dependendo da combinação desejada. O processamento dos dados é dividido em duas fases: treinamento

e teste. Na fase de treinamento para cada amostra é calculada a expansão polinomial, um exemplo do cálculo da mesma é dada pela expressão 4.2. O passo seguinte é concatenar M, para a correta concatenação deve ser utilizada a expressão 4.5. Finalizada a fase de testes o próximo passo é encontrar a solução para a equação 4.6. Para encontrar os coeficientes da função polinomial é utilizado a equação 4.9. Em seguida o algoritmo irá realizar o processamento da fase de testes, onde para cada amostra é realizada o cálculo da expressão 4.1 que representa o produto escalar do vetor solução de coeficientes polinomias obtido pela equação 4.9 com o vetor de expansão de teste conforme exemplo mostrado em 4.2 . Para finalizar o algoritmo realiza a classificação através do sinal algébrico obtido na equação 4.3.

Algorithm 1: Classificador Polinomial

Entrada: Características $\mathbf{x} = [x_1 \dots x_d]^T$, Grau n e Rótulos $\mathbf{b} = [b_1\ b_2\ \cdots b_N]$

Saída: Classificação em ω_1 ou ω_2

Dado N amostras $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$;
//Fase de Treinamento
para cada *amostra de treinamento* **hacer**
    Calcular a expansão polinomial $\mathbf{p}_n(\mathbf{x})$;
    Concatenar $\mathbf{M} = [\mathbf{p}_n(\mathbf{x}_1)\ \mathbf{p}_n(\mathbf{x}_2)\ \cdots \mathbf{p}_n(\mathbf{x}_N)]^T$;
Calcular $\mathbf{a}^* = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{b}$;
//Fase de Teste
para cada *amostra de teste* **hacer**
    Calcular $y_i = \mathbf{a}^{*T}\mathbf{p}_n(\mathbf{v})$;
    se $y_i > 0$ **então**
        $\mathbf{v} = \omega_1$;
    senão
        $\mathbf{v} = \omega_2$;

Figura 4.3: Pseudo-código ilustrando a implementação do CP.

4.3 Classificadores

De maneira a investigar a predição dos resultados das partidas de futebol, alguns dos principais algoritmos de aprendizagem de máquinas foram utilizados neste trabalho. Foram escolhidos os algoritmos Naive Bayes, árvore de decisão, MLP, RBF e SVM. A escolha desses classificadores é baseada no estado da arte pois foram utilizados em vários trabalhos tais como [34, 62, 67].

4.3.1 Naive Bayes

Naive Bayes (NB) é a implementação de um classificador probabilístico simples baseado na aplicação do teorema de Bayes que calcula um conjunto de probabilidades através da contagem da frequência e combinações de valores num dado conjunto de dados [38]. Os valores das características de cada classe são tratados de forma independente. Este fator permite a este classificador lidar de maneira eficiente com grandes quantidades de dados evitando problemas de dimensionalidade.

Todos detalhes da teoria e do funcionamento do classificador Naive Bayes pode ser encontrado em [21]. Uma ideia básica do funcionamento é apresentado em [38]. Dada uma base de dados qualquer, com um conjunto de atributos de tamanho n $(x_1, x_2, x_3, \ldots, x_n)$ e um conjunto m de classes $(C_1, C_2, C_3, \ldots C_m)$. A classificação será baseada no teorema de Bayes seguindo as seguintes equações:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4.11}$$

$P(X)$ precisa ser maximizado pois é necessário que a mesma possua igual valor para todas as classes, logo, obtemos a equação:

$$P(C_i|X) = P(X|C_i)P(C_i) \tag{4.12}$$

O Classificador NB parte da premissa de que os atributos são condicionalmente independentes, ou seja, não exite dependência entre eles. Sendo assim as atribuições de classe

das amostras de testes são baseadas nas seguintes equações:

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \tag{4.13}$$

$$\arg\max_{c_i}\{P(X|C_i)P(C_i)\} \tag{4.14}$$

Por exemplo, se uma amostra vier a ser avaliada e a probabilidade $P(C_2|X)$ tem o maior valor entre as probabilidades $P(C_k|X)$ para todas as classes k então tal amostra irá pertencer à classe C_2 de acordo com a regra de decisão do classificador NB. A estrutura geral do classificador NB está apresentado na Figura 4.4.

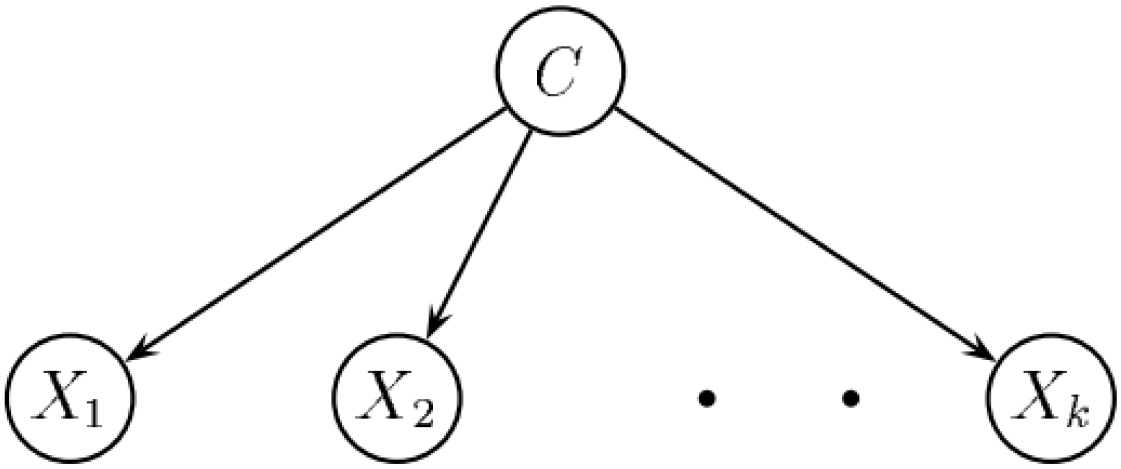


Figura 4.4: Estrutura geral do Naive Bayes.

4.3.2 Árvore de Decisão

A árvore de decisão (AD) é um algoritmo de aprendizagem que utiliza regras de classificação e uma representação baseada na estrutura de dados denominada árvore proposto em [54]. Uma árvore de decisão é composta de um conjunto de nós. Um nó pode ser de dois tipos básicos: nós de decisão, que fragmentam a decisão e são utilizados para a construção de um caminho através da árvore, e os nós terminais, folhas, que permitem determinar a qual classe a instância que está sendo avaliada.

Maiores detalhes envolvendo a teoria e os algoritmos utilizados para a implementação de classificadores baseados em árvore de decisão podem ser encontrados em [21]. A ideia básica do funcionamento da árvore de decisão é apresentada na figura 4.5.

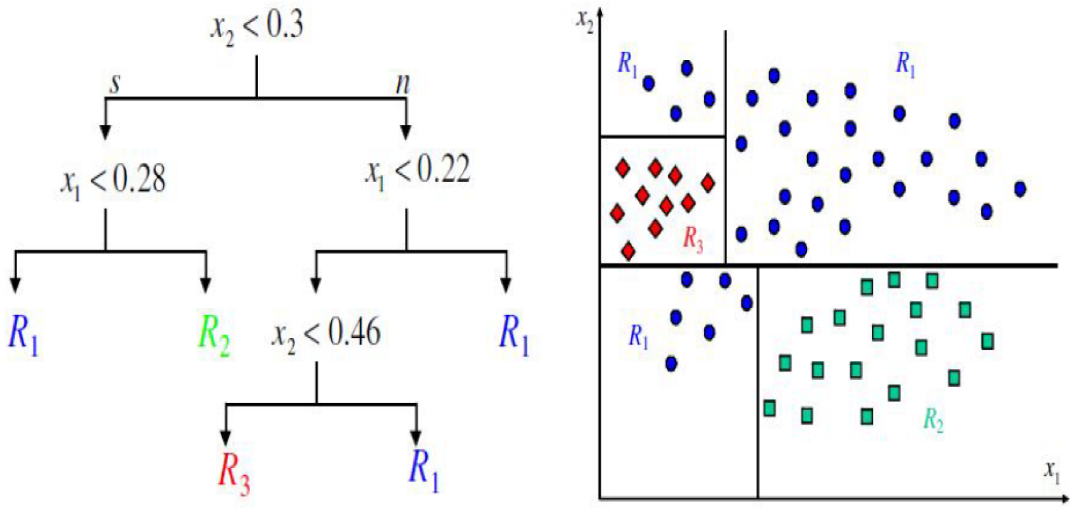


Figura 4.5: Funcionamento do classificador árvore de decisão, extraída de [21].

A Figura 4.5 apresenta um exemplo composto por três classes: R_1, R_2 e R_3 que são avaliados por duas variáveis: x_1 e x_2. Na imagem a direita temos um gráfico com instâncias das três classes projetados no plano cartesiano formado pelas duas variáveis. A distribuição dessas instâncias permite a separação das classes através de 4 retas, gerando 5 subgrupos de dados. A árvore de decisão irá realizar a classificação de uma nova instância através da estrutura apresentada a esquerda na figura. Cada reta irá dar origem a um nó de decisão onde uma tomada de decisão é realizada e cada um dos 5 subgrupos de dados irá dar origem a uma folha da árvore.

4.3.3 Multilayer Perceptron

O Multilayer Perceptron (MLP) é uma rede neural artificial com uma ou mais camadas ocultas e uma camada de saída de perceptrons originalmente proposto por [14]. Os sinais de entrada são propagados sempre para a direção à frente camada por camada. Este algoritmo utiliza uma técnica de aprendizagem supervisionada denominada *"back propagation para*

treinamento". O modelo é capaz de classificar dados não-linearmente separáveis. Todos os detalhes teóricos e práticos envolvendo essa técnica pode ser encontrados em [21].

A Figura 4.6 apresenta a estrutura básica de funcionamento do MLP, onde podemos observar a presença de uma camada de entrada que propaga o sinal para a camada escondida - poderia ser mais de uma camada escondida - que por sua vez propaga o sinal para uma camada de saída.

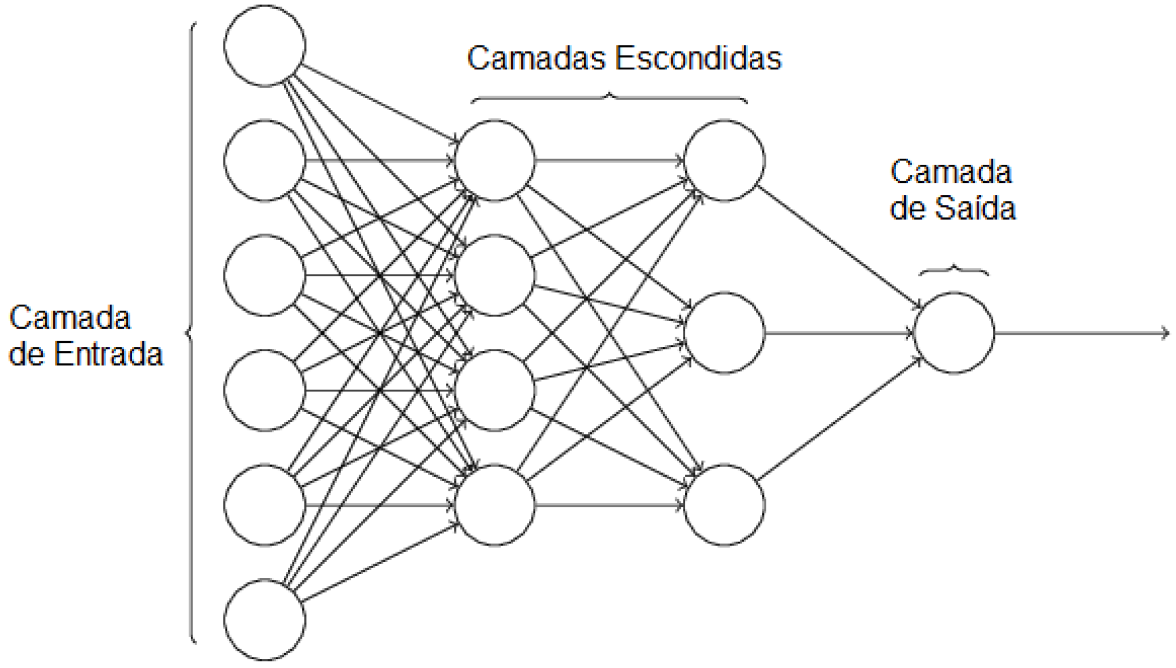


Figura 4.6: Estrutura do classificador MLP, extraída de [56].

4.3.4 Radial Basis Function

O Radial Basis Function (RBF) é uma rede neural artificial, composta de três camadas: entrada, escondida e de saída. Os nós da camada de entrada são capazes de enviar sinais para nós da camada escondida que por sua vez enviam sinais apenas para os nós da camada de saída que irão processar e realizar a decisão lógica. Esse método foi proposto por [7], onde se encontram todos os detalhes de sua correta implementação.

A Figura 4.7 apresenta a estrutura do classificador RBF. Neste algoritmo a conversão da camada de entrada para a camada oculta é não-linear e a camada oculta realiza a

transformação não linear de vetores de entrada. A camada de saída implementa uma soma ponderada das saídas da camada oculta. A entrada em uma rede RBF é não-linear enquanto a saída é linear.

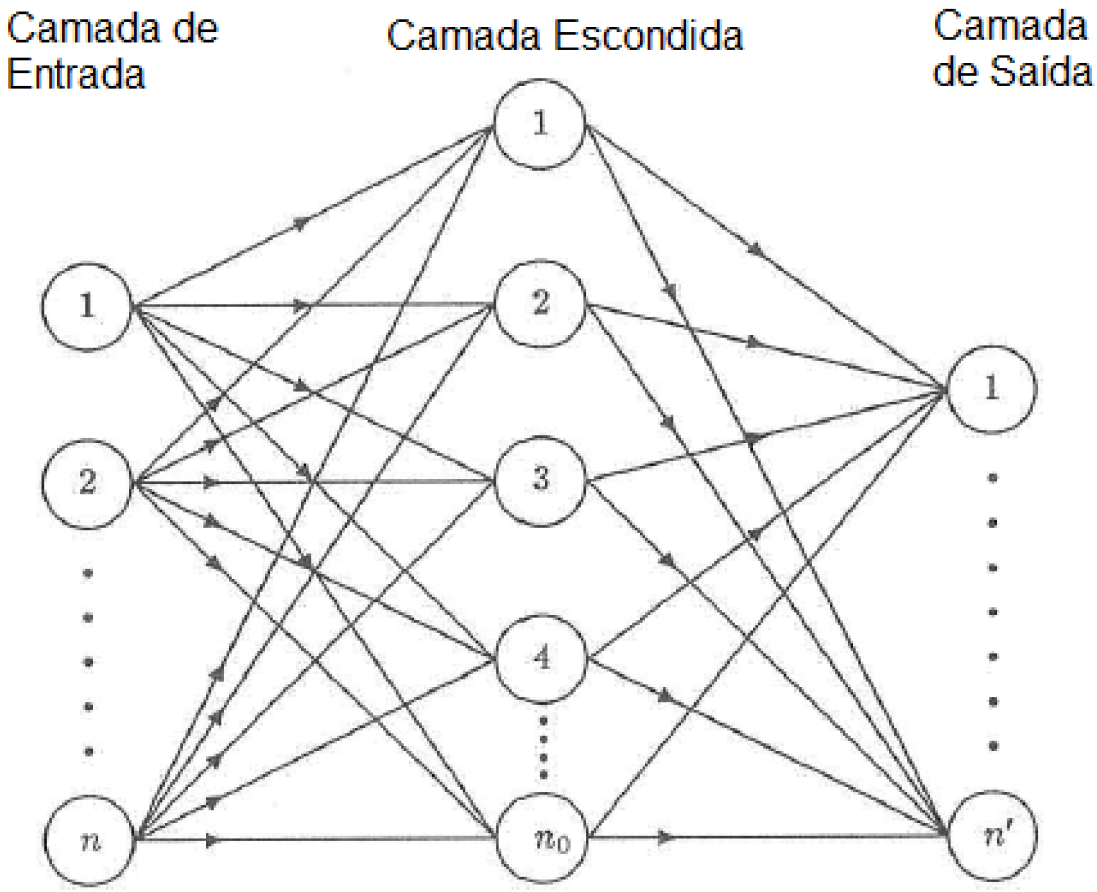


Figura 4.7: Estrutura do classificador RBF, adaptada de [7].

4.3.5 Support Vector Machine

O Support Vector Machine (SVM) é um modelo de aprendizagem utilizado em problemas de classificação de duas classes. Os recursos de entrada são mapeados para construir um espaço dimensional que é usado para elaborar uma superfície de decisão. Originalmente proposto por [68] ele tem uma excelente capacidade de classificar classes linearmente separáveis. Posteriormente [12] apresenta uma nova versão do algoritmo na qual é possível classificar as classes não lineares por meio de um espaço dimensional maior para a classificação dos

dados. Todos os detalhes para a correta implantação e a teoria envolvendo o SVM podem ser encontrado em [21]. Para o caso específico desta tese, foi utilizado o SVM com o *kernel polinomial*.

Martins [45] apresenta uma visão geral sobre o SVM que pode ser ilustrada através da Figura 4.8. O SVM é um algoritmo que constrói hiperplanos com o objetivo de encontrar hiperplanos ótimos, ou seja, hiperplanos que maximizem a margem de separação das classes, para separar os padrões de treinamento em diferentes classes. A margem é a menor distância entre os exemplos do conjunto de treinamento e o hiperplano utilizado para a separação das classes. Ela determina quão bem duas classes podem ser separadas. Os vetores suporte são realçados por círculos externos nos padrões.

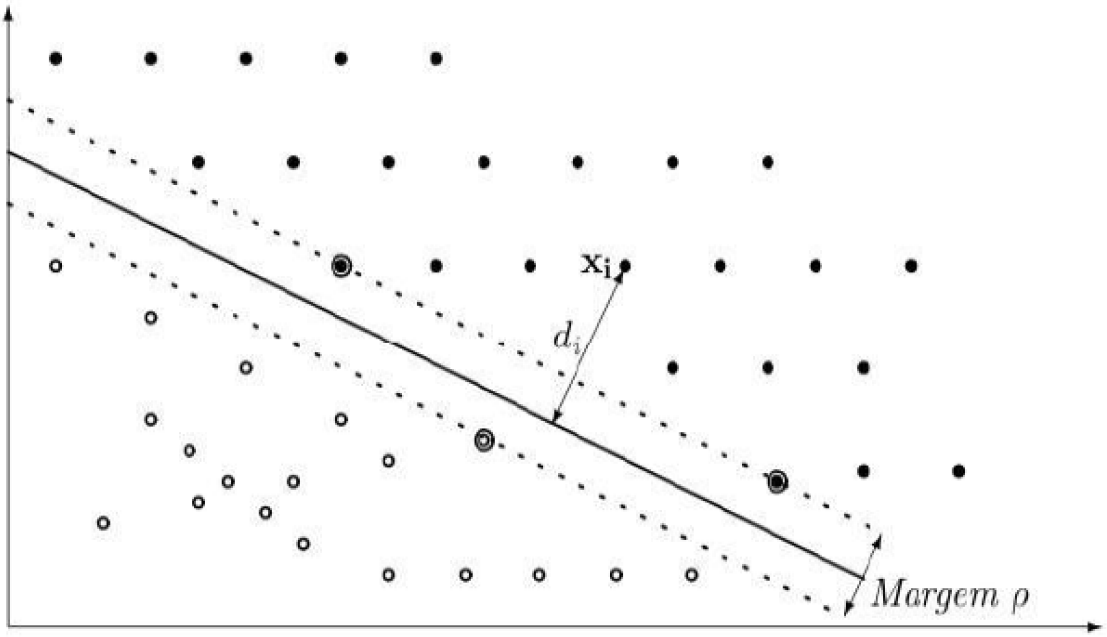


Figura 4.8: Visão geral do SVM, extraída de [45].

4.3.6 Detalhes da implementação dos Classificadores

Os classificadores utilizados no processo de investigação desta tese foram implementados através da ferramenta Waikato Environment for Knowledge Analysis (WEKA) [29]. Os

parâmetros de configuração para cada um dos algoritmos utilizados estão apresentados na tabela 4.1.

Tabela 4.1: Parâmetros de configuração para cada um dos classificadores utilizados no WEKA.

Classificadores	**Parâmetros**
NB	weka.classifiers.bayes.NaiveBayes
AD	weka.classifiers.trees.J48 -C 0.25 -M 2
MLP	weka.classifiers.functions.MultilayerPerceptron -L 0.1 -M 0.05 -N 3000 -V 0 -S 0 -E 40 -H a
RBF	weka.classifiers.functions.RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1
SVM	weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V 1 -W 1 -K

4.4 Seletores de Características

Para validar a eficácia do classificador polinomial foram realizados testes com os métodos: Análise de Componentes Principais e Relief que são técnicas utilizadas na literatura para a redução da dimensionalidade, conforme podemos observar em [62] e [66].

4.4.1 Análise de Componentes Principais

Análise de Componentes Principais ou do inglês Principal Components Analysis(PCA) proposto por [37] e [72], é uma abordagem estatística aplicada para analisar conjuntos multivariados de dados, principalmente quando é frequentemente desejável reduzir a dimensionalidade. O método é baseado na transformação ortogonal para converter dados correlacionados em um conjunto de valores de variáveis linearmente não correlacionadas que são chamados de componentes principais. O objetivo é definir os domínios de padrões inativos através do cálculo da quantidade máxima de variância obtendo menor número de componentes principais. Portanto, o número de componentes deve ser menor ou igual ao número de variáveis originais.

A Figura 4.9 apresenta a transformação PCA que reduz um grande número de variáveis (características) para um menor número de novas variáveis denominadas componentes principais. A componente principal é o arranjo que melhor representa a distribuição de dados. Já a componente secundária é perpendicular à componente principal [57]. As amostras tridimensionais são projetadas em um espaço de componente bidimensional que mantém a maior variação nos dados. O espaço de componentes aplica um subespaço linear do espaço original de alta dimensão, onde os dados se situam ou estão próximos. O PCA simplesmente rotaciona o espaço de dados original de modo que os componentes principais sejam o eixo de um novo sistema de coordenadas - o espaço de componentes. Isso pode ser estendido matematicamente para mais de três dimensões originais.

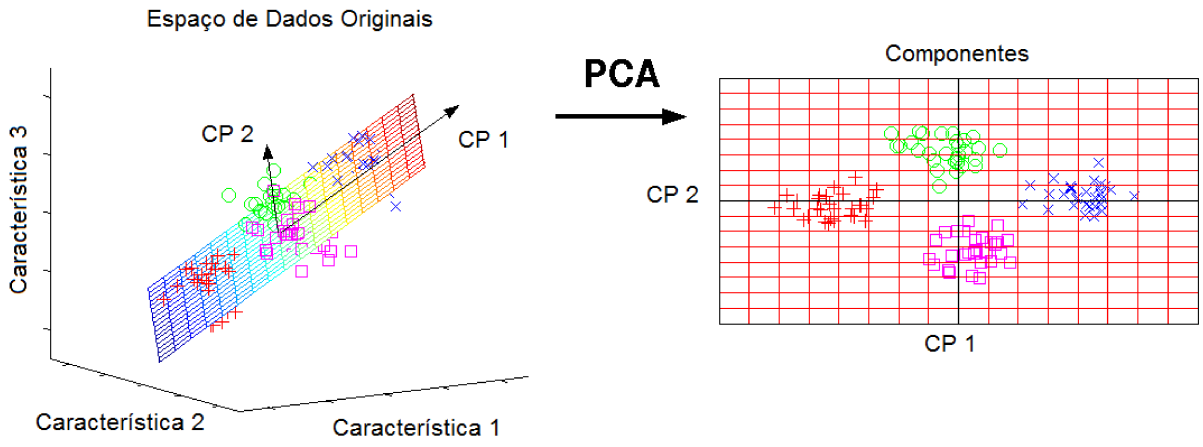


Figura 4.9: Representação gráfica do método PCA, adaptado de [57].

4.4.2 Relief

Relief é um algoritmo de seleção de caracterísiticas baseado na ideia de estimar as características de acordo com a proximidade existentes entre as instâncias avaliadas. Considerando uma dada instância, o algoritmo procura seus dois vizinhos mais próximos, um denominado de *"nearest hit"* e o outro mais distante, chamado de *"nearest miss"*. Embora este algoritmo não diferencie características redundantes e quando aplicado sobre um baixo número de instâncias de treinamento pode vir a fornecer resultados ruins, ele é tolerante ao

ruído, pode ser aplicado em dados binários ou contínuos e não é dependente de heurística [40].

O algoritmo utiliza uma função $diff$ que é responsável por retornar a diferença entre os valores de duas instâncias para cada característica. Lima [42] apresentou um exemplo do funcionamento do algoritmo considerando a busca de três "nearest hit"e três "nearest miss". A mesma pode ser visualizada em 4.10. Para um perfeito entendimento do método Relief é recomendado a leitura de [29, 39].

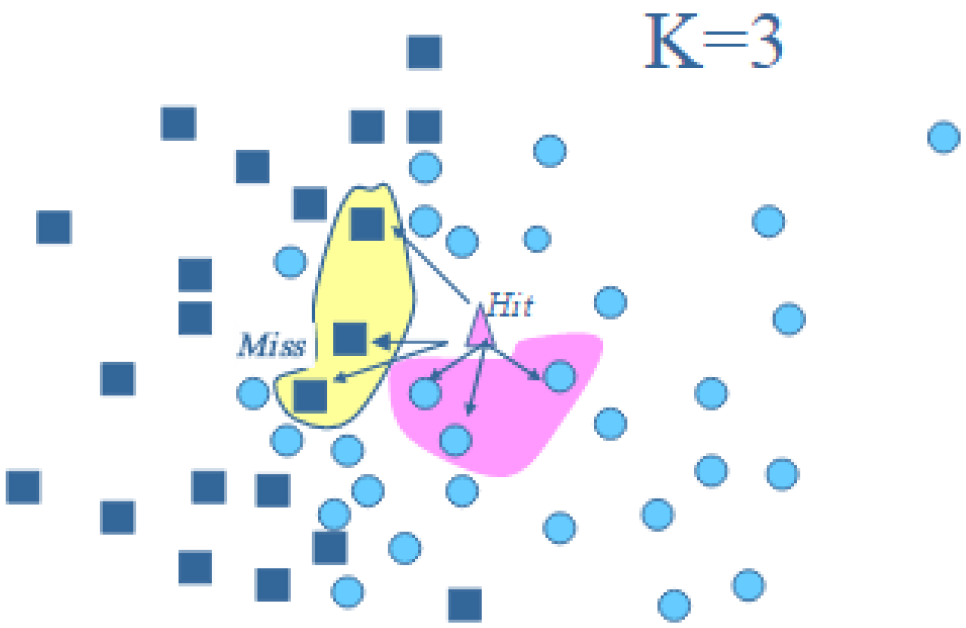


Figura 4.10: Funcionamento do algoritmo Relief, adaptado de [42].

4.4.3 Implementação dos Seletores de Características

Os seletores de características utilizados no processo de investigação desta tese também foram implementados através da ferramenta WEKA [28]. Os parâmetros de configuração para cada um dos algoritmos utilizados estão apresentados na tabela 4.2. É importante ressaltar que trata-se da configuração padrão dos dois métodos, ou seja, ambos estão livres

para realizar a seleção de características tanto no que diz respeito à quantidade bem como quanta na relevância das mesmas.

Tabela 4.2: Parâmetros de configuração para cada um dos seletores de características utilizados no WEKA.

PCA	weka.attributeSelection.PrincipalComponents
	OPTIONS
	centerData: false
	maximumAttributeNames: 5
	transformBackToOriginal: false
	varianceCovered: 0.95
	Método de Busca
	weka.attributeSelection.Ranker
	OPTIONS
	generateRanking: true
	numToSelect: -1
	threshold: -1.7976931348623157E308
Relief	weka.attributeSelection.ReliefFAttributeEval
	numNeighbours:10
	sampleSize:,-1
	seed: 1
	sigma: 2
	weightByDistance : FALSE
	Método de Busca
	weka.attributeSelection.Ranker
	OPTIONS
	generateRanking: true
	numToSelect: -1
	threshold: -1.7976931348623157E308

4.5 Considerações Finais Deste Capítulo

Este capítulo apresentou a hipótese de trabalho desta tese que é a de que o Classificador Polinomial, uma técnica amplamente utilizada como classificador de padrões, possa ser usada também como um algoritmo de seleção de características. Apresentou também outros classificadores e algoritmos de seleção de características presentes no estado da arte e que farão parte da metodologia de teste que será apresentada no capítulo 5.

Capítulo 5

Metodologia

5.1 Introdução

Este capítulo apresenta a metodologia adotada para a predição de resultados de partidas de futebol. A Seção 5.2 ilustra os fluxogramas da metodologia adotada, as bases de dados utilizadas são apresentadas na Seção 5.3, já a Seção 5.4 mostra informações a respeito do vetor de características, as técnicas de validação do método são abordadas na seção 5.5, a Seção 5.6 apresenta as métricas utilizadas para avaliar a precisão do método proposto, os testes estatísticos que comprovam a relevância das características e os resultados obtidos se encontram na seção 5.7 e finalmente as considerações finais deste capítulo são sintetizadas na seção 5.8.

5.2 Visão Geral

A primeira parte dos testes envolve a utilização de apenas classificadores para a predição de resultados de partidas de futebol. A base de dados é composta por dados obtidos através do processo de *scout* dando origem a um vetor de características. O vetor de características irá alimentar a fase de treinamento com uma partição da base de dados e servirá como parâmetro para a fase seguinte que são os testes. Ao final de cada instância testada o classificador irá realizar uma decisão lógica e classificar aquela instância, conforme ilustrado na Figura 5.1.

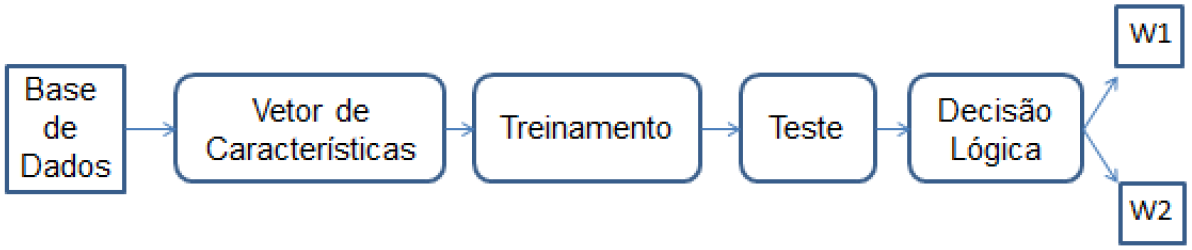


Figura 5.1: Fluxograma da metodologia sem a utilização de seletores de características.

A segunda parte dos testes envolve uma nova etapa que é a seleção de características. Nesse caso a base de dados dá origem a um novo vetor de características; que será submetido às fases de treinamento e testes. O processo encerra-se com a decisão lógica do classificador que definirá a qual classe pertence a instância submetida, conforme ilustrado na Figura 5.2.

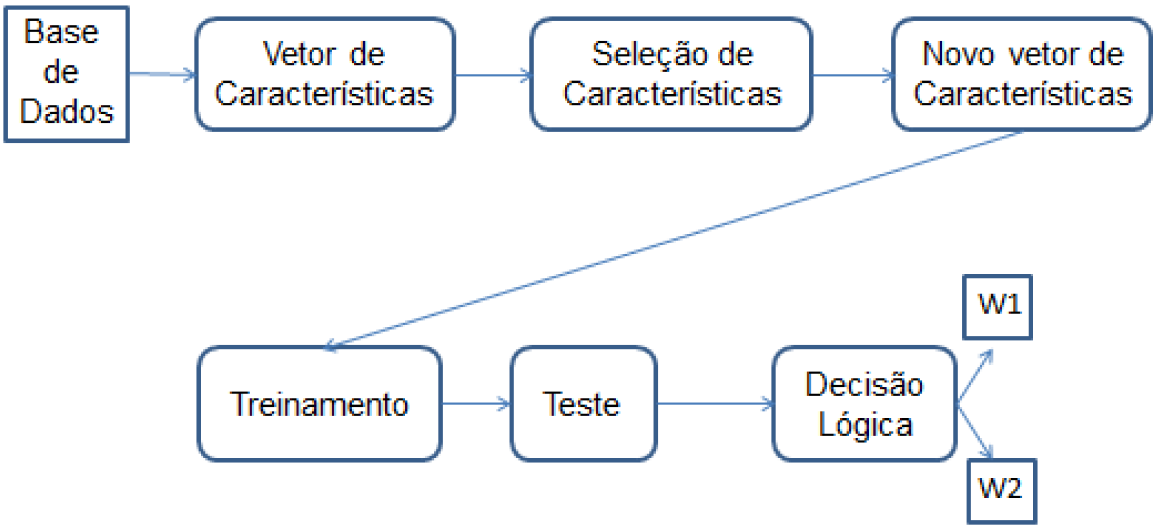


Figura 5.2: Fluxograma da metodologia com a utilização de seletores de características.

5.3 Base de dados utilizadas

Os dados utilizados para a abordagem proposta foram os resultados das partidas de futebol, obtidos nos seguintes campeonatos: campeonato inglês temporada 2014/15 (CI 2014/15), campeonato espanhol temporada 2014/15 (CE 2014/15) e campeonatos brasileiro temporadas de 2010 (CB 2010) e 2012 (CB 2012).

Campeonato Brasileiro de 2010:

O Campeonato Brasileiro de 2010 (CB 2010) foi disputado pelas seguintes equipes: Atlético Goianiense, Atlético Mineiro, Atlético Paranaense, Avaí, Botafogo, Ceará, Corinthians, Cruzeiro, Flamengo, Fluminense, Goiás, Grêmio, Grêmio Prudente, Guarani, Internacional, Palmeiras, São Paulo, Santos, Vasco da Gama e Vitória. Estas vinte equipes enfrentaram as outras dezenove duas vezes, uma na condição de mandante e outra na de visitante, totalizando assim 38 rodadas sendo disputadas dez partidas em cada uma dessas rodadas totalizando 380 partidas jogadas em todo o campeonato.

Campeonato Brasileiro de 2012:

O Campeonato Brasileiro de 2012 (CB 2012) foi disputado pelas seguintes equipes: Atlético Goianiense, Atlético Mineiro, Bahia, Botafogo, Corinthians, Coritiba, Cruzeiro, Figueirense, Flamengo, Fluminense, Grêmio, Internacional, Náutico, Palmeiras, Ponte Preta, Portuguesa, Santos, São Paulo, Sport e Vasco da Gama. Da mesma maneira que aconteceu com o CB 2010, o CB 2012 foi disputado em 38 rodadas de dez partidas cada totalizando 380 partidas jogadas em todo o campeonato.

Campeonato Inglês temporada de 2014/15:

O Campeonato Inglês temporada de 2014/15 (CI 2014/15) foi disputado pelas seguintes equipes: Arsenal, Aston Villa, Burnley, Chelsea, Crystal Palace, Everton, Hull City, Queens Park Rangers, Leicester City, Liverpool, Manchester City, Manchester United, Newcastle United, Southampton, Stoke City, Sunderland, Swansea City, Tottenham Hotspur, West Bromwich Albion e West Ham United. O formato de disputa é o mesmo do campeonato brasileiro formando assim uma base de dados de 380 jogos.

Campeonato Espanhol temporada de 2014/15:

O Campeonato Espanhol temporada de 2014/15 (CE 2014/15) foi disputado pelas seguintes equipes: Almería, Athletic Bilbao, Atlético Madrid, Barcelona, Celta de Vigo, Córdoba, Deportivo La Coruña, Eibar, Elche, Espanyol, Getafe, Granada, Levante, Málaga, Rayo Vallecano, Real Madrid, Real Sociedad, Sevilla, Valência e Villareal. A base de dados é de 380 jogos, no mesmo formato do campeonato brasileiro.

A utilização de quatro bases de dados diferentes tem como objetivo demonstrar a eficácia da solução proposta. Pois apesar do formato ser o mesmo por outro lado os times envolvidos com seus respectivos jogadores e sistemas de jogo que produziram um quantitativo de fundamentos bastante distintos.

5.4 Vetor de Características

Os vetores de características utilizados para solucionar o problema estão diretamente ligadas ao processo de coleta de dados que foi apresentado no capítulo 3. Os 27 fundamentos ali apresentados irão compor parcial ou integralmente o vetor de características das quatro bases. Ressalta-se que não existe uma padronização dos fundamentos que serão coletados pelo processo de *scout* mas que a definição de um fundamento não sofrerá alteração.

Os dados referentes ao CB 2010 e ao CB 2012 foram coletados pela empresa Match Report com sede na cidade de Campinas através do projeto scoutonline da linha de inovação tecnológica financiado pela Fapesp. Os dados foram organizados de acordo com os resultados da partida sob a ótica da equipe mandante (vitória, empate e derrota). A metodologia de coleta leva em consideração os 27 fundamentos apresentados no capítulo 3. Sendo assim o vetor de características possui 54 variáveis, os 27 fundamentos executados pelas equipes mandantes e visitantes.

O vetor de características dos CB 2010 e CB 2012 é definido como: (x_1) assistência do time da casa, (x_2) assistência do time visitante, (x_3) bola recebida time da casa, (x_4) bola recebida time visitante, (x_5) bola recuperada time da casa, (x_6) bola recuperada time visitante, (x_7) bola perdida time da casa, (x_8) bola perdida time visitante, (x_9) cartão amarelo time da casa, (x_{10}) cartão amarelo time visitante, (x_{11}) cartão vermelho time da casa,

(x_{12}) cartão vermelho time visitante, (x_{13}) cruzamento certo time da casa, (x_{14}) cruzamento certo time visitante, (x_{15}) cruzamento errado time da casa, (x_{16}) cruzamento errado time visitante, (x_{17}) defesa time da casa, (x_{18}) defesa time visitante, (x_{19}) desarme completo time da casa, (x_{20}) desarme completo time visitante, (x_{21}) desarme incompleto time da casa, (x_{22}) desarme incompleto time visitante, (x_{23}) drible certo time da casa, (x_{24}) drible certo time visitante, (x_{25}) drible errado time da casa, (x_{26}) drible errado time visitante, (x_{27}) escanteio cedido time da casa, (x_{28}) escanteio cedido time visitante, (x_{29}) escanteio conquistado time da casa, (x_{30}) escanteio conquistado time visitante, (x_{31}) falta recebida time da casa, (x_{32}) falta recebida time visitante, (x_{33}) falta cometida time da casa, (x_{34}) falta cometida time visitante, (x_{35}) finalização certa time da casa, (x_{36}) finalização certa time visitante, (x_{37}) finalização errada time da casa, (x_{38}) finalização errada time visitante, (x_{39}) total finalização time da casa, (x_{40}) total finalização time visitante, (x_{41}) gol time da casa, (x_{42}) gol time visitante, (x_{43}) gol contra time da casa, (x_{44}) gol contra time visitante, (x_{45}) impedimento time da casa, (x_{46}) impedimento time visitante, (x_{47}) lançamento certo time da casa, (x_{48}) lançamento certo time visitante, (x_{49}) lançamento errado time da casa, (x_{50}) lançamento errado time visitante, (x_{51}) passe certo time da casa, (x_{52}) passe certo time visitante, (x_{53}) passe errado time da casa e (x_{54}) passe errado time visitante.

Os dados referentes aos CI 2014/15 e CE 2014/15 pertencem a uma base pública e está disponível em http://www.football-data.co.uk/. A exemplo do que ocorre com as bases dos campeonatos brasileiros os dados foram organizados de acordo com o resultado obtido pela equipe mandante (vitória, empate e derrota). Essas bases de dados são objetos de outros estudos que compõe o estado da arte como podemos observar em [62, 67]. A metodologia de coleta leva em consideração 9 dos fundamentos apresentados no capítulo 3, gerando assim um vetor de características com 18 variáveis.

O vetor de características para o caso dos CI 2014/15 e CE 2014/15 é definido da seguinte maneira: (x_1) gols time da casa, (x_2) gols time visitante, (x_3) total de finalizações time da casa, (x_4) total de finalizações time visitante, (x_5) finalização certa time da casa, (x_6) finalização certa time visitante, (x_7) finalização errada time da casa, (x_8) finalização errada time visitante, (x_9) escanteio conquistado time da casa, (x_{10}) escanteio conquistado

time visitante, (x_{11}) falta cometida time da casa, (x_{12}) falta cometida time visitante, (x_{13}) impedimento time da casa, (x_{14}) impedimento time visitante, (x_{15}) cartão amarelo time da casa, (x_{16}) cartão amarelo time visitante, (x_{17}) cartão vermelho time da casa e (x_{18}) cartão vermelho time visitante.

Estes vetores de características descritos pertencem à primeira parte dos testes realizados conforme apresentado na figura 5.1. Na segunda parte dos testes realizados exitem dois vetores de características, conforme é possível observar na figura 5.2, o primeiro vetor de características que é exatamente esse que foi apresentado até aqui, o novo vetor de características é construído após a aplicação dos algoritmos de seleção das características selecionando as características mais relevantes para a classificação dos dados envolvidos.

5.5 Técnicas de Validação do Método

As técnicas de validação do método utilizadas são: cross validation, apresentado na Seção 5.5.1 e sliding window apresentado na Seção 5.5.2 e os idetalhes da Implementação das técnicas de validação do método na Seção 5.5.3. Tais técnicas foram utilizadas para calcular o desempenho dos algoritmos utilizados para predição dos resultados das partidas de suas respectivas base de dados. As três possíveis saídas de resultados das partidas são: vitória, empate e derrota, que deu origem a três grupos de investigação: vitória x empate, vitória x derrota e empate x derrota.

5.5.1 Cross Validation

A técnica cross validation ou técnica de validação cruzada consiste em uma divisão dos dados em n subgrupos de tamanho igual onde (n-1) subgrupos são utilizados para treinamento e 1 subgrupo é utilizado para teste. Todos os n subgrupos são testados cada um em uma iteração. Para o caso específico desta tese foi utilizada a técnica 10-folds-cross-validation proposto por [8]. Os jogos foram distribuídos em 10 subconjuntos, depois um dos 10 subconjuntos é escolhido como teste e os restantes 9 subconjuntos como treinamento. A precisão do modelo é então calculado para este 1x9 subconjuntos. Este procedimento é

repetido para cada um dos 10 subconjuntos, e assim terminamos com 10 valores de precisão. A precisão final é calculada como uma média de 10 valores, conforme é possível observar na figura 5.3.

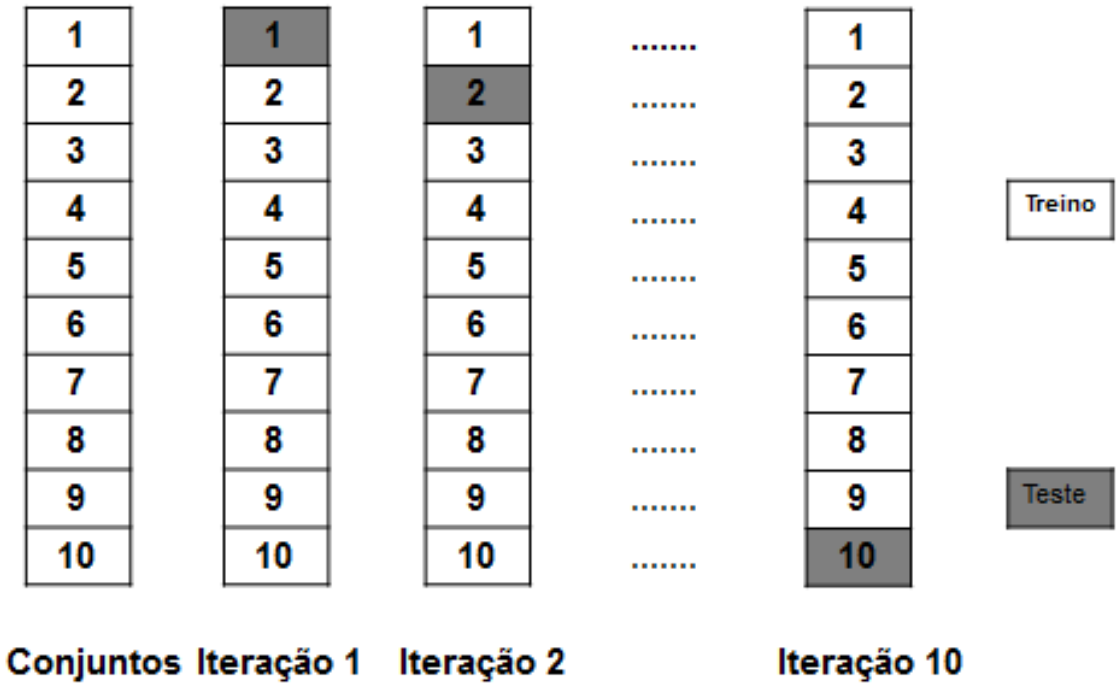


Figura 5.3: Funcionamento da técnica cross validation.

5.5.2 Slinding Window

A técnica slinding window ou técnica da janela deslizante também foi utilizada para estimativa da precisão dos algoritmos testados. A escolha por tal técnica devesse ao fato de que o fator tempo é considerado essencial na resolução de problemas de predição de resultados segundo [69]. A técnica cria duas janelas sobre subconjuntos de dados uma janela de treinamento e uma janela de teste. Para esse caso em específico cada rodada do campeonato é visto como um subconjunto de dados. O tamanho da janela de treinamento é 4 e o tamanho da janela de teste é 2, este procedimento foi adotado para os dados das 38 rodadas do campeonato, que resultou em 17 conjuntos de dados para avaliação . A precisão final é calculada como uma média da precisão dos 17 conjuntos. O primeiro conjunto de

dados é formado pela janela de treinamento $1, 2, 3, 4$ e a janela de teste $5, 6$. O segundo conjunto de dados é formado pela janela de treinamento $3, 4, 5, 6$ e a janela de teste $7, 8$. e assim sucessivamente até chegarmos ao último conjunto de dados é formado pela janela de treinamento $33, 34, 35, 36$ e a janela de teste $37, 38$, conforme é possível observar na figura 5.4.

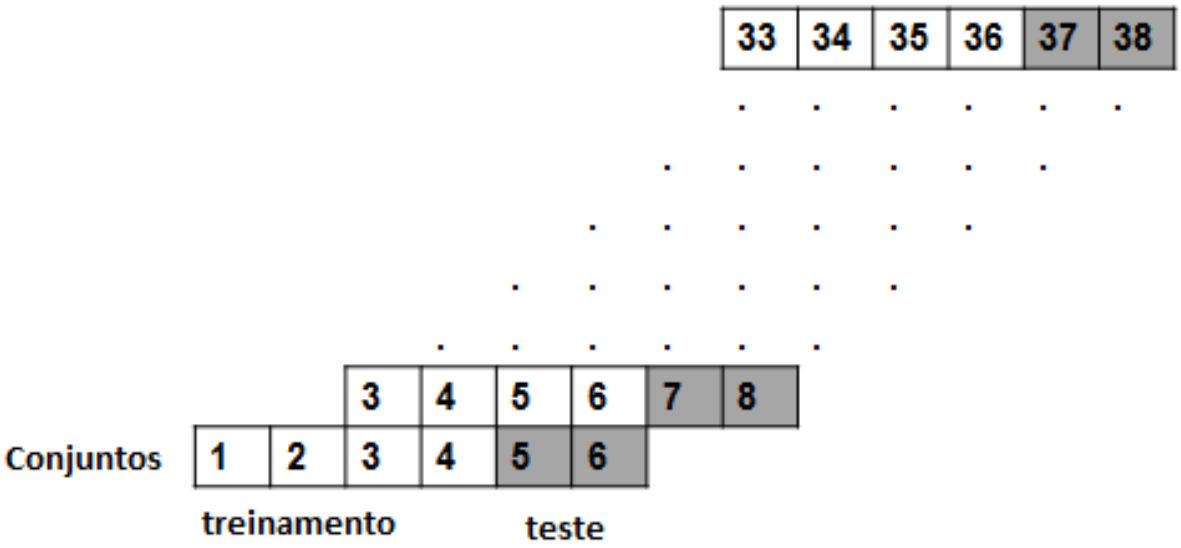


Figura 5.4: Funcionamento da técnica sliding window.

5.5.3 Implementação das técnicas de validação do método

As técnicas de validação do método foram implementadas através do software MATLAB R2012b. No caso da técnica *cross validation* a implementação se deu através da função *crossvalind* com o parâmetro *Kfold* que irá dividir o conjunto em diferentes partições geradas aleatoriamente. No caso da técnica *slinding window* a divisão dos conjuntos é realizada através da descrição apresentada na subseção anterior.

5.6 Métrica Utilizada

Para avaliar a abordagem proposta a métrica de precisão acurácia (AC) é utilizada. Tal métrica é bastante utilizada no estado da arte como podemos observar em [22], [19].

A acurácia é definida como a proporção de previsões corretas relacionadas ao número de amostras avaliadas:

$$AC = \frac{(VP + VN)}{(P + N)}$$

onde (VP) é a taxa de casos verdadeiros positivos , (VN) é a taxa de verdadeiros negativos ambos calculados em relação a todos os positivos (P) e todos os negativos (N).

5.7 Testes Estatísticos

Esta seção descreve o conceito do teste de *T-Student*, na subseção 5.7.1 utilizado aqui para avaliar a relevância das características escolhidas para a solução do problema. Apresenta também o *Rank de Friedman* para classificar e avaliar a relevância dos classificadores envolvidos, detalhado na subseção 5.7.2.

5.7.1 Teste T-Student

Esta subseção descreve o procedimento estatístico para teste de hipóteses, que é um procedimento bastante padrão comumente usado por pesquisadores para testar alguma afirmativa. Em estatística, a hipótese é uma afirmativa sobre uma propriedade da população e o teste de hipótese (ou teste de significância) é um procedimento padrão para testar uma afirmativa sobre uma propriedade da população [65]. Os componentes de um teste de hipótese são:

- Hipótese nula (H_0): é uma afirmativa de que o valor de um parâmetro populacional (como padrão, média ou desvio-padrão) é igual a algum valor especificado.
- Hipótese alternativa (H_1): é a afirmativa de que o parâmetro tem um valor de que difere da hipótese nula.

A hipótese nula é testada diretamente no sentido de que se supõe que ela seja verdadeira e chegasse a uma conclusão para rejeitar H_0 ou deixar de rejeitar H_0. A estatística de teste é um valor calculado a partir dos dados amostrais e é usada para se tomar a decisão sobre a rejeição da hipótese nula. Ela é encontrada pela conversão da estatística amostral em um

escore com a suposição de que a hipótese nula seja verdadeira, portanto, pode ser usada para determinar se há evidência significativa contra a hipótese nula. A estatística de teste de média esta representada na equação 5.1 pode se basear na distribuição normal ou na distribuição t de *Student*, dependendo das condições que sejam satisfeitas [65].

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{ou} \qquad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{5.1}$$

A determinação da rejeição ou não da hipótese nula é determinada pelo valor de p, onde valor p é a probabilidade de se obter um valor da estatística de teste que seja no mínimo tão extremo quanto o que representa os dados amostrais. No caso da hipótese nula seja verdadeira. O p-valor é menor que o nível de significância proposto (α), então Z_{obs} está na região crítica e portanto, rejeitamos a hipótese nula H_0 . Por outro lado, se o $p - valor$ é maior que o nível de significância, a hipótese nula não é rejeitada, conforme podemos observar na figura 5.5. Além disso, quanto menor for o p-valor, mais "distante"estamos da hipótese nula H_0. O $p - valor$ para o ponto amostral x é definido matematicamente como

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta[W(X) \geq W(x)],$$

em que θ é um parâmetro pertencente ao espaço paramétrico Θ sob a hipótese nula (H_0) [26]. Para o caso desta tese foi seguido o critério de rejeitar a hipótese nula se o valor p for muito pequeno, tal como 0,05 ou menos como em [2].

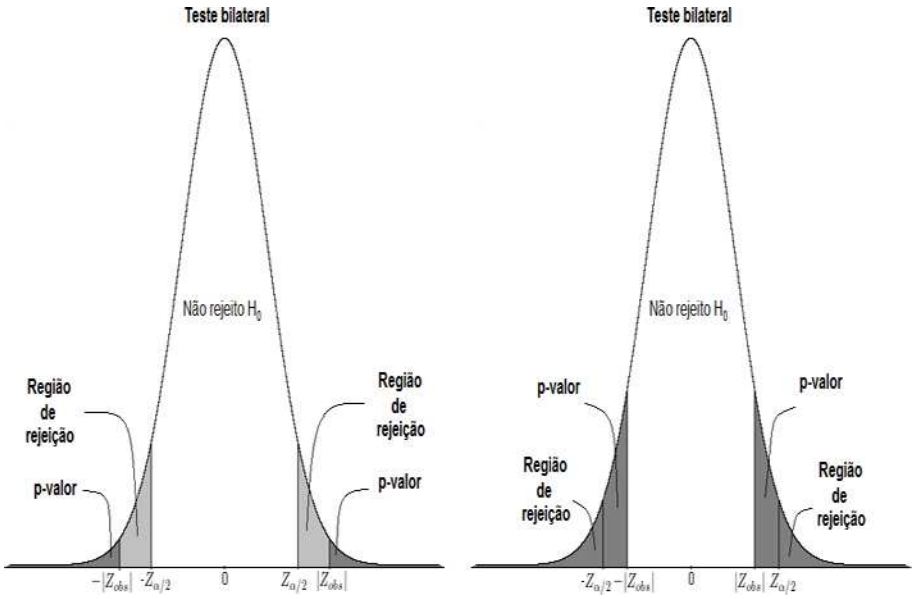


Figura 5.5: Interpretação do P Valor, adaptada de [26].

5.7.2 Rank de Friedman

O teste de Friedman é uma alternativa não paramétrica para o teste de experimentos em blocos ao acaso (RBD - Randon Blocks Design) na ANOVA regular. Este teste utiliza os ranks dos dados ao invés de seus valores brutos para o cálculo da estatística de teste [15].

O cálculo da estatística do teste de Friedman é realizado através de uma sequência simples de passos, o primeiro é ordenar as k observações da menor para a maior de forma separada em cada um dos b blocos e atribuímos os ranks 1, 2, ..., k para cada bloco da tabela de observações. Assim, a posição esperada de qualquer observação sob H_0 é (k + 1)/2. Sendo r(Xij) o rank da observação Xij definimos a soma de todos os ranks da coluna j (ou seja, de cada tratamento) por

$$R_j = \sum_{i=1}^{b} r(X_{ij}), \quad 1 \leq j \leq k.$$

Se H_0 é verdadeira, o valor esperado de Rj é E(Rj)=b(k+1)/2. Desta forma, a estatística

$$\sum_{j=1}^{k} \left(R_j - \frac{b(k+1)}{2} \right)^2$$

é uma forma intuitiva para revelar as diferenças entre os tratamentos.

A estatística do teste de Friedman será dada por

$$S = \frac{12b}{k(k+1)} \sum_{j=1}^{k} \left(\frac{R_j}{b} - \frac{k+1}{2} \right)^2 = \left[\frac{12}{bk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3b(k+1)$$

Se $F_j(t) = F(t + \tau_j)$ é a função de distribuição do tratamento j, com $j = 1, 2, ..., k$, no teste de Friedman estamos interessados em testar a hipótese H_0: $\tau_1 = \tau_2 = ... = \tau_k$ contra a hipótese alternativa de que $\tau_1 = \tau_2 = ... = \tau_k$ não são todos iguais. Neste caso, ao nível de significância α, rejeitamos a hipóteseH_0 se $S \geq s_\alpha$, caso contrário não rejeitamos a hipótese nula, em que a constante s_α é escolhida de modo que a probabilidade de erro do tipo I seja igual a α.

5.8 Considerações Finais deste capítulo

Este capítulo apresentou a metodologia proposta que tem como objetivos demonstrar que:

- A abordagem proposta é capaz de identificar resultados de partidas de futebol.
- É possível aplicar essa abordagem para investigar base de dados presentes no estado da arte.
- O classificador polinomial pode ser usado com eficiência como algoritmo de seleção de características.
- As características e os resultados obtidos pelos classificadores possuem relevância estatística.

Capítulo 6

Resultados Obtidos

6.1 Introdução

Este capítulo apresenta os resultados obtidos através dos testes realizados com a metodologia proposta no capítulo 5. A Seção 6.2 mostra os resultados obtidos pelos classificadores sem a utilização dos seletores de características. A Seção 6.3 exibe os resultados obtidos com o uso de seletores de características. A Seção 6.4 apresenta os resultados dos campeonatos brasileiros com as mesmas características das bases dos campeonatos europeus. A Seção 6.5 apresenta o resultado para os testes estatísticos. A Seção 6.6 apresenta a comparação com o estado da arte. As considerações finais são apresentadas na seção 6.7.

6.2 Resultados obtidos pelos classificadores sem a utilização de seletores de características

Esta seção apresenta os resultados obtidos pelos classificadores sem a utilização de seletores de características. Os resultados obtidos utilizando a amostragem Cross Validation estão na subseção 6.2.1, enquanto os resultados obtidos pela técnica Slinding Window se encontram na subseção 6.2.2.

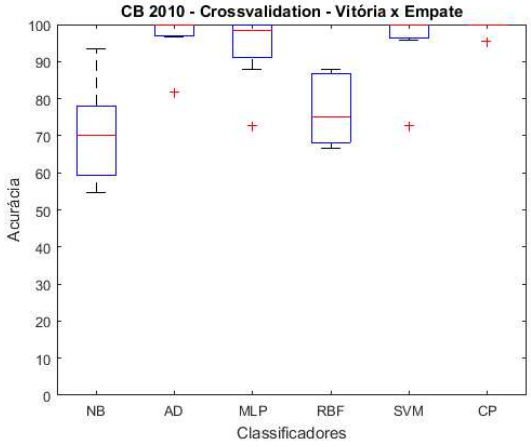
6.2.1 Resultados com o Cross Validation

A técnica de validação do método cross validation foi aplicada para os dados dos campeonatos brasileiro de 2010 e 2012, espanhol temporada 2014/15 e inglês 2014/15. Os dados foram divididos em 3 casos nas seguintes combinações: vitória x empate, vitória x derrota e empate x derrota. Os resultados serão apresentadas através de duas abordagens: uma tabela contendo a média e o desvio padrão, obtidos nos dez conjuntos de dados testados para cada caso e uma representação gráfica utilizando *boxplot*.

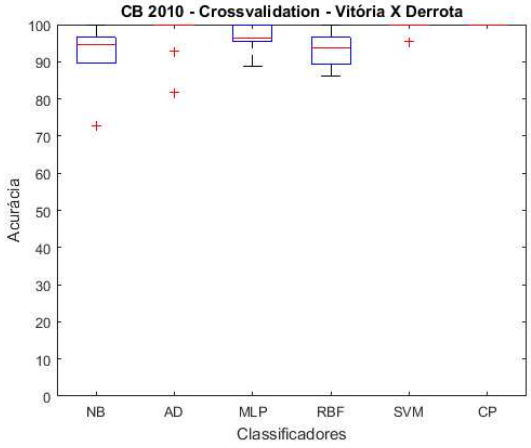
No caso do campeonato brasileiro de 2010 o CP obteve como acurácia os seguintes valores: 99,5±1,4 para a combinação vitória x empate, 100,0±00 para vitória x derrota e 99,3±1,9 no caso empate x derrota. O SVM apresentou acurácia de 96,1±8,4 na combinação vitória x empate, 99,5±1,4 empate x derrota e de 97,2±4,1 para empate x derrota. O RBF obteve como resultados: 76,5±9,2, 93,1±4,6 e 78,4±7,1. O classificador MLP apresentou a performance de: 94,4±8,7, 96,2±3,5 e 94,0±5,0 . Já a abordagem AD obteve taxas de acurácia: 97,2±5,6, 97,4±5,9, 96,7±3,7. Enquanto no algoritmo NB a acurácia obtida foi de: 69,7±12,0, 92,6±7,9 e 74,7±5,6. Tais valores estão representados na tabela 6.1. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.1.

Tabela 6.1: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010.

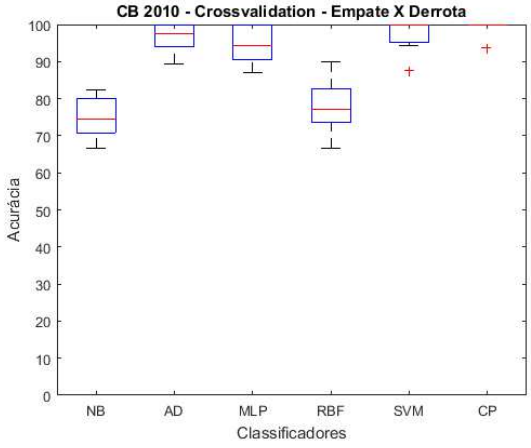
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	69,7±12,0	92,6±7,9	74,7±5,6
AD	97,2±5,6	97,4±5,9	96,7±3,7
MLP	94,4±8,7	96,2±3,5	94,0±5,0
RBF	76,5±9,2	93,1±4,6	78,4±7,1
SVM	96,1±8,4	99,5±1,4	97,2±4,1
CP	99,5±1,4	100,0±00	99,3±1,9



(a)



(b)



(c)

Figura 6.1: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010.

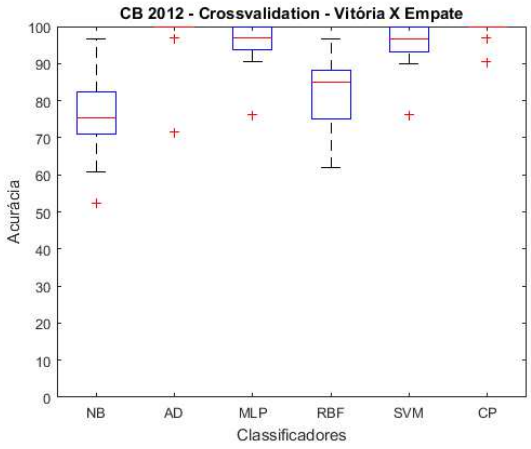
Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.1, obtidas para a combinação vitória x empate o CP obteve os melhores resultados com 99,5±1,4, em seguida vieram AD com 97,2±5,6, SVM com 96,1±8,4 , MLP com 94,4±8,7, RBF com 76,5±9,2 e NB com 69,7±12,0. Tal tendência é confirmada pelos valores plotados no gráfico 6.1a, onde ocorre uma junção dos primeiros e terceiros quartis bem como os limites inferiores e superiores em torno da mediana, com apenas um valor discrepante e mesmo assim bem próximo a mediana. Os classificadores AD e SVM possuem gráficos bem similares no que diz respeito a mediana, os quartis e os limites, a diferença pequena entre as duas médias é explicada justamente nos valores discrepantes, onde um dos conjuntos de dados do SVM ficou bem abaixo dos demais. O MLP vem na sequência já com uma mediana um pouco inferior e com um primeiro quartil e um limite inferior mais alongado. O desempenho abaixo dos classificadores RBF e NB também podem ser observados. O RBF obteve uma mediana superior ao NB, além disso existe uma distância menor entre os limites inferior e superior para os resultados obtidos quando se compara o desempenho de ambos. Ao se analisar a combinação vitória x derrota é possível afirmar novamente que o CP obteve o melhor desempenho com 100,0±00, seguido por SVM com 99,5±1,4, AD com 97,4±5,9, MLP com 96,2±3,5, RBF com 93,1±4,6 e NB com 92,6±7,9, sendo que esses dois últimos tiveram uma melhora significativa. Através do gráfico 6.1b é possível visualizar que o CP obteve a mediana ideal e que o resultados foi obtido em todos os conjuntos. É possível entender ainda o motivo da inversão do desempenho dos classificadores AD e SVM, onde ambos tiveram a mesma mediana, os mesmos quartis, mas o SVM teve um valor discrepante enquanto a AD obteve dois, sendo que em um deles o valor ficou bem abaixo do restante. Já os classificadores RBF e NB obtiveram resultados bem superiores nessa combinação mas as posições foram mantidas muito em virtude de um valor discrepante do NB. O MLP também melhorou seu desempenho muito em função de não existirem valores discrepantes para esse caso. Para a combinação empate x derrota a mesma tendência pode ser observada tanto com as médias e seus respectivos desvio padrões onde temos por ordem de desempenho: CP 99,3±1,9, SVM 97,2±4,1, AD 96,7±3,7, MLP 94,0±5,0, RBF 78,4±7,1 e NB 74,7±5,6. A mesma tendência é confirmada pela análise do gráfico 6.1c onde a mediana e os quartis se

concentram para o caso do CP com apenas um valor discrepante, o SVM apesar de possuir um valor discrepante se destaca pela mediana e também pela proximidade com a mesma tanto o quartil quanto do limite inferior, a AD se destaca em relação ao MLP tanto em relação a mediana tanto quanto ao limite inferior, sendo que ambos não possuem valores discrepantes. RBF e NB tem desempenho superior para o RBF, muito em função dos quartis e limites superiores.

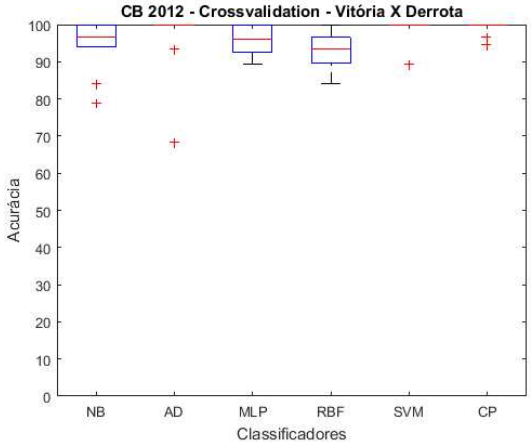
Com a base de dados do campeonato brasileiro de 2012 o CP obteve acurácias de 98,7±3,0, 99,1 ±1,8 e 99,0±3,1 nas combinações de vitória x empate, vitória x derrota e empate x derrota respectivamente. O classificador SVM apresentou taxas de acertos de: 94,8±7,3, 98,9±3,3 e 98,1±3,3 . Por sua vez a técnica RBF alcançou taxas de: 81,1±11,7, 92,9±5,2 e 78,9±9,9 . MLP obteve acurácias de: 95,0±7,3, 95,6±3,7 e 95,0±5,0 . Os resultados do algoritmo AD foi de: 96,8±8,9 , 96,1±9,9 e 98,0±4,7 , enquanto que o NB conseguiu como resultado: 75,0±12,3, 94,2±7,1 e 77,0±5,7 . A tabela 6.2 apresenta os resultados obtidos. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.2.

Tabela 6.2: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012.

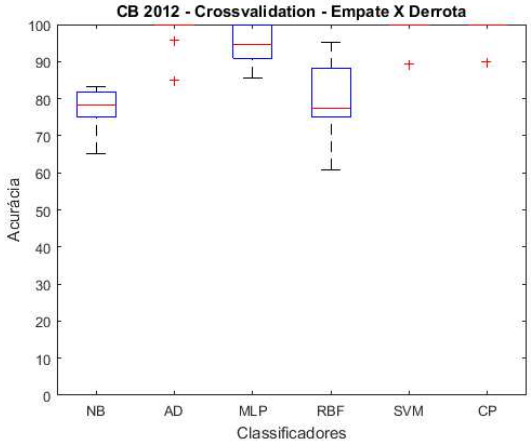
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	75,0±12,3	94,2±7,1	77,0±5,7
AD	96,8±8,9	96,1±9,9	98,0±4,7
MLP	95,0±7,3	95,6±3,7	95,0±5,0
RBF	81,1±11,7	92,9±5,2	78,9±9,9
SVM	94,8±7,3	98,9±3,3	98,1±3,3
CP	98,7±3,0	99,1 ±1,8	99,0±3,1



(a)



(b)



(c)

Figura 6.2: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012.

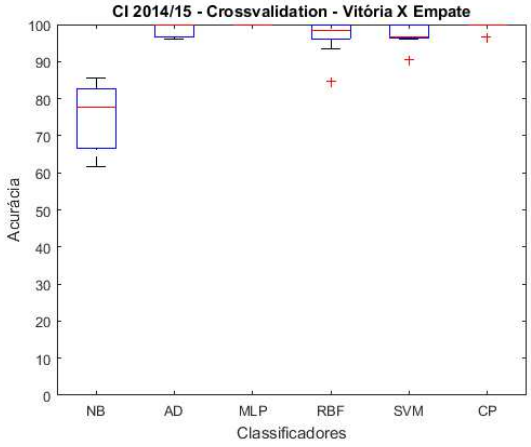
Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.2, obtidas para a combinação vitória x empate o CP obteve os melhores resultados com 98,7±3,0, seguido por AD com 96,8±8,9, MLP com 95,0±7,3 , SVM com 94,8±7,3, RBF com 81,1±11,7 e NB com 75,0±12,3. Através do gráfico 6.2a é possível visualizar que a maioria dos valores obtidos pelo CP está situado em torno da mediana da mesma maneira que ocorre com a AD, ambos possuem também apenas dois valores discrepantes mas um deles, para o caso da AD ficou bem abaixo. MLP e SVM também possuem desempenhos bastante próximos um do outro, com para o MLP no quartil e limite inferior. Os dois classificadores com pior desempenho foram o RBF e o NB com os quartis e o limite superior sendo os responsáveis pelo melhor desempenho do RBF. Para a combinação vitória x derrota temos a seguinte ordem de acurácia média: CP(99,1 ±1,8), SVM(98,9±3,3), AD(96,1±9,9), MLP(95,6±3,7), NB(94,2±7,1) e RBF(92,9±5,2). Através do gráfico 6.2b podemos observar que CP, SVM e AD possuem boa parte dos valores em torno da mediana, o CP possui apenas dois valores discrepantes mais bem próximos a mediana, já o SVM possui apenas um valor discrepante mais com desempenho inferior a 90%, enquanto que a AD possui dois valores discrepantes sendo que um deles com desempenho inferior a 70% o que explica a média mais abaixo que a dos outros dois. Apesar da mediana do MLP ser inferior a do NB, o mesmo possui dois valores discrepantes bem abaixo do limite inferior do MLP, enquanto no caso do RBF a mediana ficou abaixo das dos demais classificadores, não existindo registros de valores discrepantes. No caso da combinação empate x derrota temos por ordem de desempenho: CP(99,0±3,1), SVM(98,1±3,3), AD(98,0±4,7), MLP(95,0±5,0), RBF(78,9±9,9) e NB(77,0±5,7). Através do gráfico mostrado na figura 6.2c os classificadores CP, SVM e AD possuem a mediana ideal, com CP e SVM com apenas dois valores discrepantes, sendo que o do último fica um pouco abaixo do que o primeiro, para o caso da AD temos dois valores discrepantes. Na sequência temos o MLP com a quarta melhor mediana e com limite inferior superior ao do RBF e do NB. Apesar da mediana mais baixa do RBF em comparação com o NB, seu desempenho é compensado com maiores quartis e limites superiores.

A base do campeonato inglês 2014/15 apresentou as seguintes acurácias: CP(99,6±1,0,

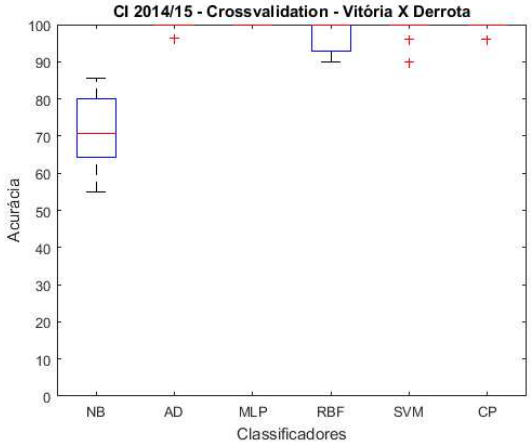
99,6±1,0 e 99,4±1,6), SVM(97,2±2,9, 99,0±1,0 e 98,6±3,2), RBF(96,7±4,8, 98,6±2,4 e 96,8±4,3), MLP(100,0±00, 100,0±00 e 100,0±00), AD(98,8±1,7, 99,7±0,9 e 99,6±1,1) e NB(75,1±8,7, 95,7±3,5 e 71,2±9,3) para as combinações vitória x empate, vitória x derrota e empate x derrota respectivamente. Tais valores são apresentados na tabela 6.3. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.3.

Tabela 6.3: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Inglês temporada de 2014/15.

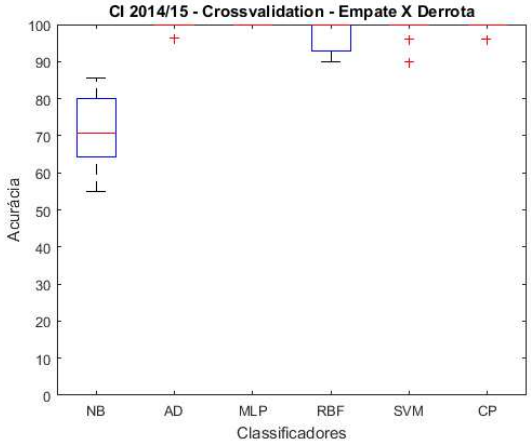
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	75,1±8,7	95,7±3,5	71,2±9,3
AD	98,8±1,7	99,7±0,9	99,6±1,1
MLP	100,0±00	100,0±00	100,0±00
RBF	96,7±4,8	98,6±2,4	96,8±4,3
SVM	97,2±2,9	99,0±1,0	98,6±3,2
CP	99,6±1,0	99,6±1,0	99,4±1,6



(a)



(b)



(c)

Figura 6.3: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Inglês temporada de 2014/15..

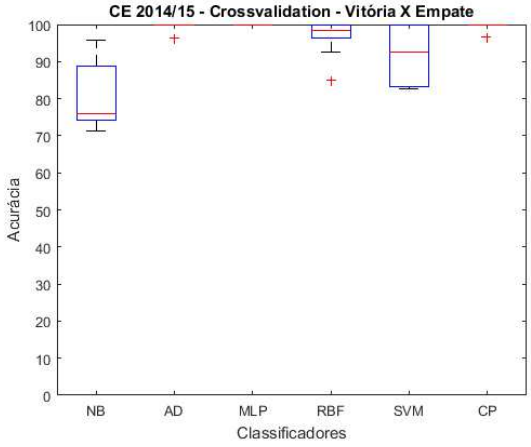
Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.3, obtidas para a combinação vitória x empate é possível perceber que a ordem de desempenho foi: MLP(100,0±00), CP(99,6±1,0), AD(98,8±1,7), SVM(97,2±2,9), RBF(96,7±4,8) e NB(75,1±8,7). Através do gráfico 6.3a podemos observar que o MLP possui todos os valores em cima da mediana ideal, o CP apesar da mediana ideal possui apenas um valor discrepante, a AD possui mediana ideal mas com quartil e limite inferior um pouco abaixo da mesma. SVM e RBF possuem desempenhos similares, com vantagem para o SVM pois apesar da mediana mais abaixo do que a do RBF seu quartil e limites inferiores estão apenas um pouco abaixo da mediana, além disso ambos possuem um valor discrepante sendo o do RBF bem abaixo do que o do SVM. O classificador NB completa o conjunto com valores inferiores aos demais. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: MLP(100,0±00), CP(99,6 ±1,1), AD(99,7±0,9), SVM(99,0±1,0), RBF(98,6±2,4) e NB(95,7±3,5). Através do gráfico 6.3b podemos observar que MLP e CP mantêm o comportamento do gráfico anterior. AD e SVM possuem melhora na mediana mas os valores discrepantes impedem um desempenho melhor na comparação com o MLP e o CP. O RBF consegue melhorar sua mediana com uma ligeira perda nos quartis e no limite inferior mas eliminando valores discrepantes. Já no caso do NB houve queda na mediana e também nos quartis inferiores, superiores bem como nos seus limites. No caso da combinação empate x derrota as melhores acurácias obtidas foram: MLP(100,0±00), AD(99,6±1,1), CP(99,4±1,6), SVM(98,6±3,2), RBF(96,8±4,3) e NB(71,2±9,3). Através do gráfico 6.3c podemos observar que mais uma vez o MLP conseguiu todos os valores junto a mediana ideal. AD e CP mantêm o comportamento da combinação anterior, entretanto para esse caso a AD leva vantagem por conta do melhor desempenho do valor discrepante. Para os demais não houve alteração de desempenho em relação ao gráfico anterior.

Para os dados referentes ao campeonato espanhol 2014/15 novamente os resultados obtidos são bastante significativos e estão representados na tabela 6.4. O CP obteve as acurácias de: 99,6±1,0, 99,6±1,0 e 99,6±1,2. O SVM teve como resutados: 92,3±7,4, 100,0±00 e 88,0±7,6. Para o RBF obtivemos: 96,6±4,8, 100,0±00 e 95,3±5,4. O MLP conseguiu taxas de 100,0±00, 100,0±00 e 100,0±00. A abordagem AD alcançou: 99,6±1,1, 99,2±1,6 e

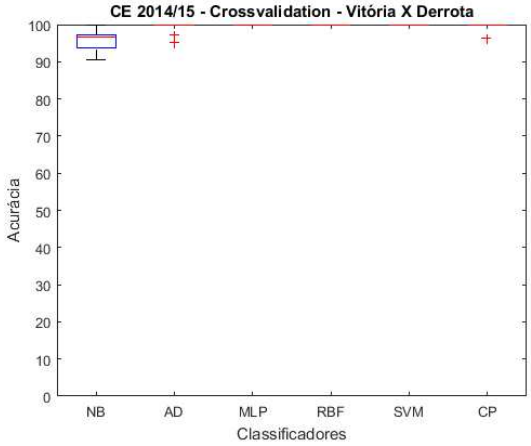
99,5±1,4 . Enquanto NB obteve: 80,7±8,6, 95,9±2,9 e 72,7±10,3, nas combinações vitória x empate, vitória x derrota e empate x derrota respectivamente. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.4.

Tabela 6.4: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

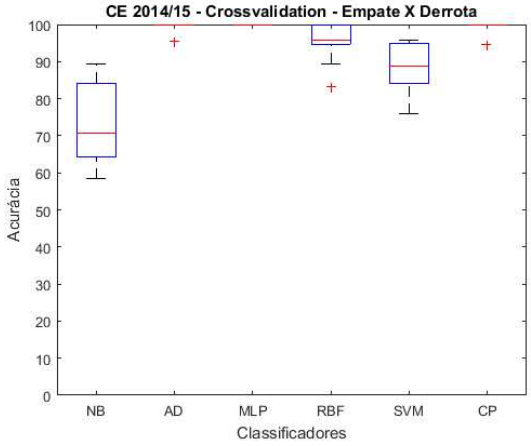
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	80,7±8,6	95,9±2,9	72,7±10,3
AD	99,6±1,1	99,2±1,6	99,5±1,4
MLP	100,0±00	100,0±00	100,0±00
RBF	96,6±4,8	100,0±00	95,3±5,4
SVM	92,3±7,4	100,0±00	88,0±7,6
CP	99,6±1,0	99,6±1,0	99,6±1,2



(a)



(b)



(c)

Figura 6.4: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.4, obtidas para a combinação vitória x empate é possível perceber que a ordem de desempenho foi: MLP(100,0±00), CP(99,6±1,0), AD(99,6±1,1), RBF(96,6±4,8), SVM(92,3±7,4) e NB(80,7±8,6). Através do gráfico 6.4a podemos observar que o MLP possui todos os valores juntos a mediana ideal, apesar da mediana ideal CP e AD possuem um valor discrepante cada, com desempenhos idênticos. Na sequência temos em ordem de desempenho RBF, SVM e NB, com diferenças de comportamento na mediana e nos quartis e limites tanto superiores quanto inferiores.Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: MLP, RBF e SVM com (100,0±00), CP(99,6 ±1,1), AD(99,2±1,6) e NB(95,9±2,9). Através do gráfico 6.4b podemos observar que o MLP, RBF e SVM possuem todos os valores juntos a mediana ideal, CP e AD também atingem a mediana ideal mas com um valor discrepante para o CP e dois para a AD, destoa desse cenário apenas NB, mas também com desempenho bem acima quando comparado a combinação anterior. No caso da combinação empate x derrota as melhores acurácias obtidas foram: MLP(100,0±00), CP(99,6±1,2), AD(99,5±1,4), RBF(95,3±5,4), SVM(88,0±7,6) e NB(72,7±10,3). Através do gráfico 6.4c podemos observar que MLP, CP e AD atingem a mediana ideal, mas com CP e AD com um valor discrepante cada enquanto o mesmo não ocorre com o MLP. É possível ainda comprovar que o desempenho de RBF, SVM e NB vem bem abaixo dos três anteriores e com os valores das medianas, quartis e limites bem diferentes uns dos outros.

6.2.2 Resultados com o Slinding Window

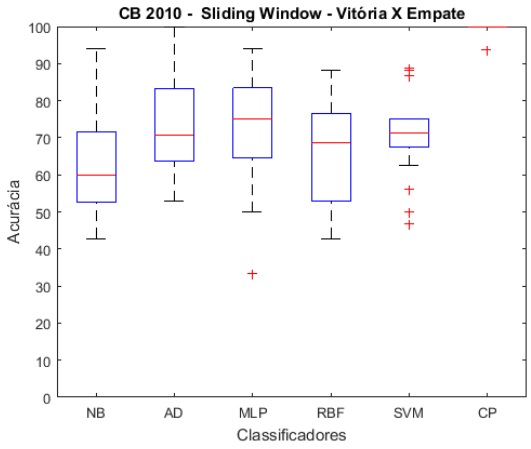
A técnica de validação do método slinding window também foi aplicada para os dados dos quatro campeonatos em estudo: brasileiro de 2010 e de 2012, espanhol temporada 2014/15 e inglês 2014/15. Os dados foram para os 4 casos divididos para as seguintes combinações: vitória x empate, vitória x derrota e empate x derrota. Os resultados serão apresentadas utilizando se duas abordagens uma tabela contendo a média e o desvio padrão, obtidos nos dezessete conjuntos de dados testados para cada caso e uma representação gráfica utilizando

boxplot.

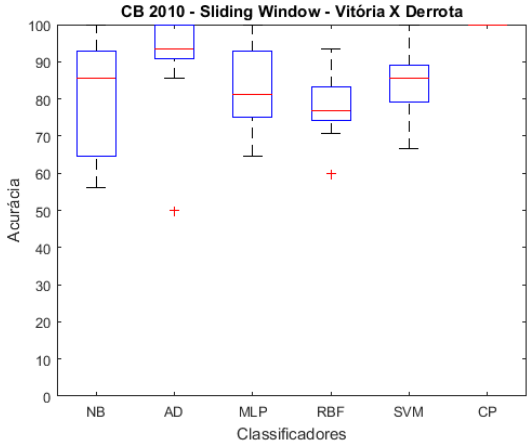
Os resultados obtidos utilizando a base de dados do campeonato brasileiro de 2010 foram de: 99,6±1,5, 100,0±00 e 100,0±00 com a utilização do CP, para o classificador SVM as acurácias obtidas foram: 70,4±11,8, 84,4±10,1 e 65,8±17,2, com o classificador RBF os valores foram: 65,6±13,8, 77,3±8,8 e 61,8±16,4, por sua vez MLP conseguiu alcançar: 72,7±15,7, 83,7±10,8 e 65,6±18,2, enquanto que AD obteve: 73,1±12,8, 92,3±11,9 e 61,7±14,2 e NB: 61,9±13,4, 81,6±15,0 e 57,6±17,2. Conforme podemos observar na tabela 6.5. Tais valores estão representados na tabela 6.1. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.5.

Tabela 6.5: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

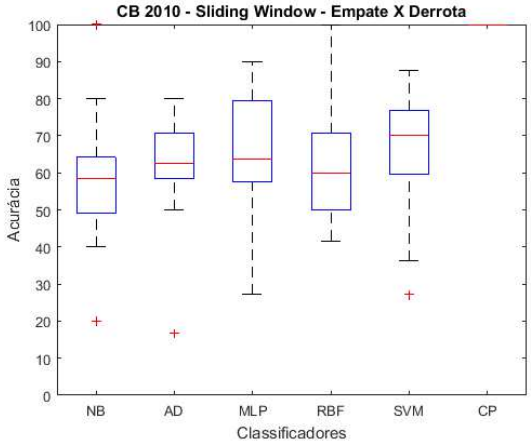
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	61,9±13,4	81,6±15,0	57,6±17,2
AD	73,1±12,8	92,3±11,9	61,7±14,2
MLP	72,7±15,7	83,7±10,8	65,6±18,2
RBF	65,6±13,8	77,3±8,8	61,8±16,4
SVM	70,4±11,8	84,4±10,1	65,8±17,2
CP	99,6±1,5	100,0±00	100,0±00



(a)



(b)



(c)

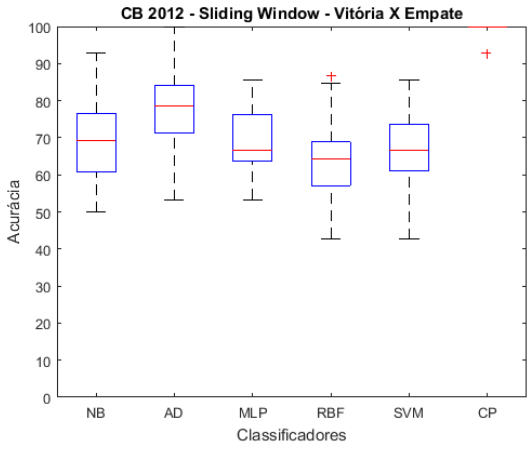
Figura 6.5: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.5, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(99,6±1,5), AD(73,1±12,8), MLP(72,7±15,7), SVM(70,4±11,8), RBF(65,6±13,8) e NB(61,9±13,4). Através do gráfico 6.5a temos que o CP atingiu a mediana ideal com apenas um valor discrepante, o segundo melhor desempenho foi o de AD muito em razão do limite superior ter atingido o valor ideal e o limite inferior ficar acima do mesmo em comparação aos demais classificadores, com desempenhos parecidos com esse temos ainda MLP e SVM, sendo que o desempenho do primeiro foi bastante prejudicado por um valor discrepante bem abaixo dos demais e o segundo com três valores discrepantes tanto abaixo quanto a acima. Um pouco abaixo desse desempenho temos os classificadores: RBF e NB. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(100,0±00), AD(92,3±11,9), SVM(84,4±10,1), MLP(83,7±10,8), NB(81,6±15,0) e RBF(77,3±8,8). Através do gráfico 6.5a podemos observar que a ordem de desempenho foi CP, AD, SVM, MLP, NB e RBF com performances bem distintas em relação a mediana, quartis e limites superiores e inferiores. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(100,0±00), SVM(65,8±17,2), MLP(65,6±18,2), RBF(61,8±16,4), AD(61,7±14,2) e NB(57,6±17,2). Através do gráfico 6.5c podemos observar que a ordem de desempenho foi CP, SVM, MLP, RBF, AD e NB com performances bem distintas em relação a mediana, quartis e limites superiores e inferiores.

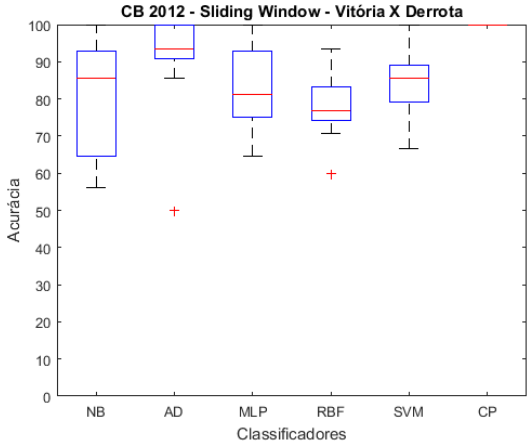
Considerando a base de dados do campeonato brasileiro de 2012 os resultados obtidos com o CP foram: 99,5±1,7, 100,0±00 e 100,0±00. O SVM alcançou 67,0±9,8, 84,4±10,1 e 64,3±13,0. Já no RBF as acurácias foram de: 64,5±10,6, 77,3±8,8 e 50,4±11,8. No MLP foram de: 69,1±9,4, 83,7±10,8 e 62,4±15,3, no AD: 78,2±11,3, 92,3±11,9 e 59,2±15,7. No NB: 69,5±11,4, 81,6±15,0 e 48,6±13,7. Conforme é possível observar na tabela 6.6. Tais valores estão representados na tabela 6.1. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.6.

Tabela 6.6: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012.

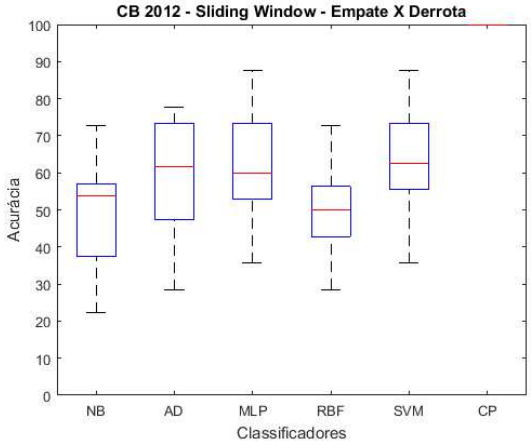
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	69,5±11,4	81,6±15,0	48,6±13,7
AD	78,2±11,3	92,3±11,9	59,2±15,7
MLP	69,1±9,4	83,7±10,8	62,4±15,3
RBF	64,5±10,6	77,3±8,8	50,4±11,8
SVM	67,0±9,8	84,4±10,1	64,3±13,0
CP	99,5±1,7	100,0±00	100,0±00



(a)



(b)



(c)

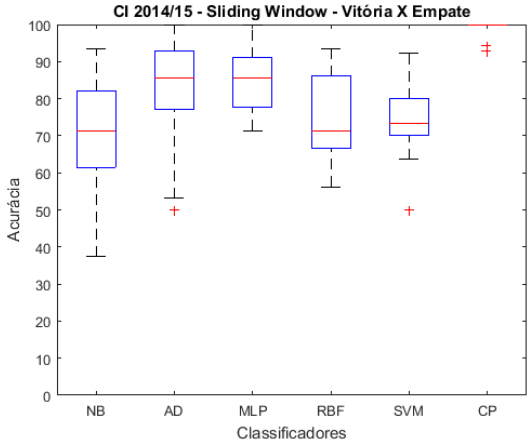
Figura 6.6: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.6, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(99,5±1,7), AD(78,2±11,3), NB(69,5±11,4), MLP(69,1±9,4), SVM(67,0±9,8) e RBF(64,5±10,6). Através do gráfico 6.6a podemos observar que a ordem de desempenho foi CP, AD, NB, MLP, SVM e RBF com performances bem distintas em relação a mediana, quartis e limites superiores e inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(100,0±00), AD(92,3±11,9), SVM(84,4±10,1), MLP(83,7±10,8), NB(81,6±15,0) e RBF(77,3±8,8). Através do gráfico 6.6a podemos observar que a ordem de desempenho foi CP, AD, SVM, MLP, NB e RBF com performances bem distintas em relação a mediana, quartis e limites superiores e inferiores. A mudança em relação a combinação anterior deve se ao fato de existirem para esse caso valores discrepantes em AD e RBF. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(100,0±00), SVM(64,3±13,0), MLP(62,4±15,3), AD(59,2±15,7), RBF(50,4±11,8) e NB(48,6±13,7). Através do gráfico 6.6c podemos observar que a ordem de desempenho foi CP, SVM, MLP, AD, RBF e NB com performances bem distintas em relação a mediana, quartis e limites superiores e inferiores.

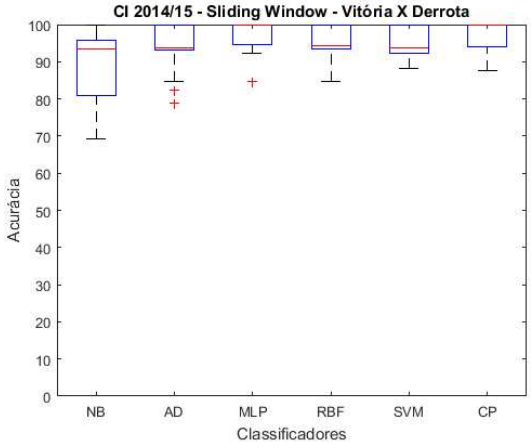
Com a base de dados do campeonato inglês de 2014/15 as acurácias alcançadas pelo CP foram de 99,2±2,1, 96,2±3,9 e 97,7±4,1. O SVM obteve: 74,8±10,6, 95,2±4,4 e 67,8±16,7. O RBF: 76,6±11,5, 95,3±4,7 e 62,4±10,6. O MLP alcançou: 85,2±8,9, 97,6±4,3 e 80,6±14,0. A AD conseguiu: 82,5±15,2, 93,8±6,4 e 80,6±16,3. E por sua vez o NB obteve: 69,7±17,2, 89,2±9,5 e 52,2±19,5. Tais valores são apresentados na tabela 6.7. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.7.

Tabela 6.7: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Inglês temporada de 2014/15.

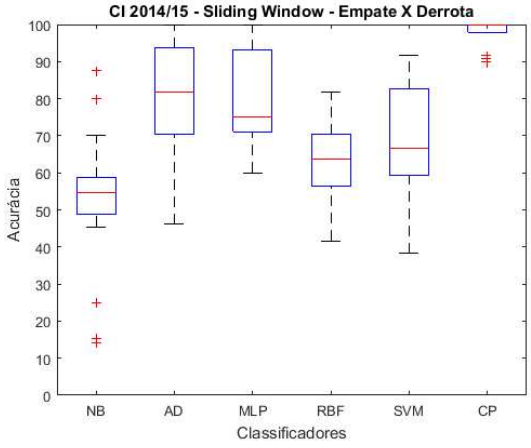
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	69,7±17,2	89,2±9,5	52,2±19,5
AD	82,5±15,2	93,8±6,4	80,6±16,3
MLP	85,2±8,9	97,6±4,3	80,6±14,0
RBF	76,6±11,5	95,3±4,7	62,4±10,6
SVM	74,8±10,6	95,2±4,4	67,8±16,7
CP	99,2±2,1	96,2±3,9	97,7±4,1



(a)



(b)



(c)

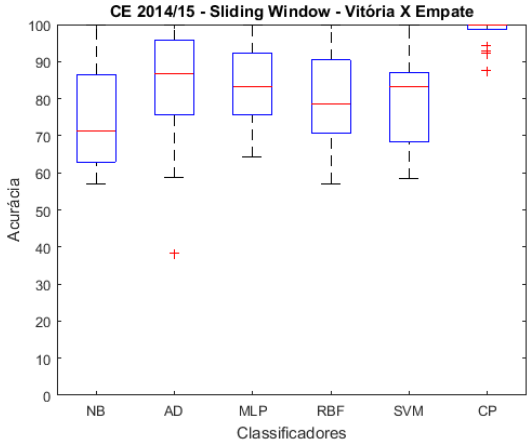
Figura 6.7: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Inglês temporada de 2014/15.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.7, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(99,2±2,1), MLP(85,2±8,9), AD(82,5±15,2), RBF(76,6±11,5), SVM(74,8±10,6) e NB(69,7±17,2). Através do gráfico 6.7a podemos observar que a ordem de desempenho foi CP, MLP, AD, RBF, SVM e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: MLP(97,6±4,3), CP(96,2±3,9), RBF(95,3±4,7), SVM(95,2±4,4), AD(93,8±6,4) e NB(89,2±9,5). Através do gráfico 6.7a podemos observar que a ordem de desempenho foi MLP, CP, RBF, SVM, AD e NB com desempenho bem superior a combinação anterior. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(97,7±4,1), MLP(80,6±14,0), AD(80,6±16,3), SVM(67,8±16,7), RBF(62,4±10,6) e NB(52,2±19,5). Através do gráfico 6.7a podemos observar que a ordem de desempenho foi CP, MLP, AD, SVM, RBF e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores.

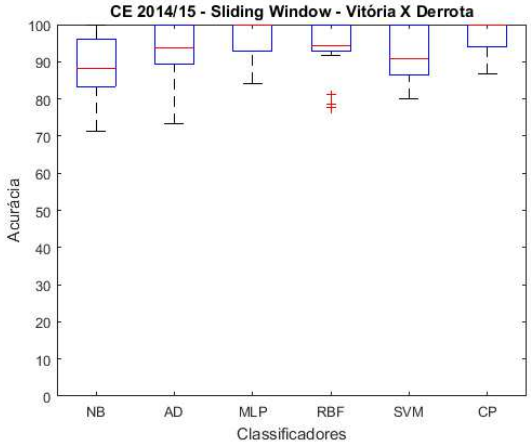
A tabela 6.8 mostra os resultados obtidos com a base de dados do campeonato espanhol de 2014/15. O CP atingiu índices de acurácia de: 98,0±3,8, 96,7±4,0 e 97,7±4,3. Por sua vez o SVM conseguiu as seguintes taxas de acertos: 78,2±12,2, 91,1±7,0 e 66,4±16,4 . Para o caso do RBF as acurácias foram de: 78,9±13,1, 93,3±7,4 e 65,7±14,8. O MLP obteve: 82,2±10,8, 95,9±5,0 e 75,6±11,2. AD com: 83,5±17,4, 92,8±7,5 e 82,2±14,2 e NB com: 75,9±14,1, 88,5±8,8 e 60,9±18,0. Também é possível observar esses resultados no gráfico 6.8.

Tabela 6.8: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

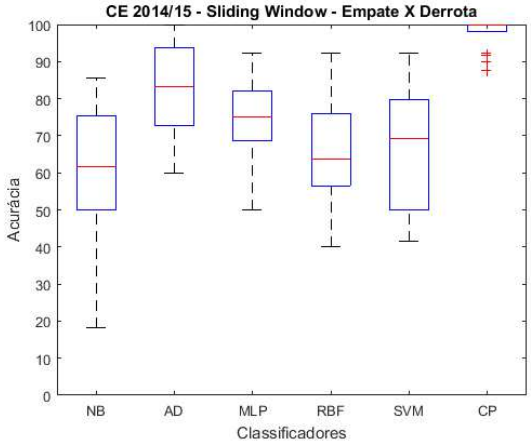
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	75,9±14,1	88,5±8,8	60,9±18,0
AD	83,5±17,4	92,8±7,5	82,2±14,2
MLP	82,2±10,8	95,9±5,0	75,6±11,2
RBF	78,9±13,1	93,3±7,4	65,7±14,8
SVM	78,2±12,2	91,1±7,0	66,4±16,4
CP	98,0±3,8	96,7±4,0	97,7±4,3



(a)



(b)



(c)

Figura 6.8: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.8, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(98,0±3,8), AD(83,5±17,4), MLP(82,2±10,8), RBF(78,9±13,1), SVM(78,2±12,2) e NB(75,9±14,1). Através do gráfico 6.8a podemos observar que a ordem de desempenho foi CP, MLP, AD, RBF, SVM e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(96,7±4,0), MLP(95,9±5,0), RBF(93,3±7,4), AD(92,8±7,5), SVM(91,1±7,0) e NB(88,5±8,8). Através do gráfico 6.8b podemos observar que a ordem de desempenho foi CP, MLP, RBF, AD, SVM e NB com performance superior a encontrada na combinação anterior. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(97,7±4,3), AD(82,2±14,2), MLP(75,6±11,2), SVM(66,4±16,4), RBF(65,7±14,8) e NB(60,9±18,0). Através do gráfico 6.8a podemos observar que a ordem de desempenho foi CP, MLP, AD, RBF, SVM e NB com performances parecidas com a combinação vitória x empate.

6.3 Resultados obtidos pelos classificadores com a utilização de seletores de características

Esta seção apresenta os resultados obtidos pelos classificadores com a utilização dos seguinte seletores de características: PCA, Relief e classifcador polinomial. Os resultados aqui serão apresentados de três maneiras distintas: uma tabela contendo a média da acurácia das três combinações (vitória x empate, vitória x derrota, empate x derrota), uma tabela contendo a varição da acurácia, diferença entre o valor base e o valor obtido por cada um dos três seletores de características(CP, PCA e Relief), e uma figura contendo um gráfico de barras com variação da acurácia distribuída pelos classificadores. Os resultados obtidos utilizando a técnica Cross Validation estão na subseção 6.3.1, enquanto os resultados obtidos pela técnica Slinding Window se encontram na subseção 6.3.2.

6.3.1 Resultados com o Cross Validation

A técnica *cross validation* foi aplicada novamente nas 4 base de dados anteriores: Campeonato Brasileiro de 2010 e 2012, Campeonato Inglês 2014/15 e Campeonato Espanhol 2014/15. Foi obtida uma média considerando as 3 combinações possíveis: vitória x empate, vitória x derrota e empate x derrota. As médias obtidas foram comparadas observando a variação em cada um dos casos.

No campeonato brasileiro de 2010 a acurácia média obtida sem a utilização de seletor e considerando o classificador NB foi de 79,0. Com a utilização do seletor Relief a acurácia se manteve a mesma, com o PCA houve perda de 10,2 e com o CP um ganho de 3,6. No caso do classificador AD a acurácia obtida sem seleção de características foi de 97,1. Com a utilização do Relief a acurácia mais uma vez se manteve a mesma, com o PCA houve uma queda de rendimento de 26.6 e com o CP ocorreu um ganho de 1,1. Quando o classificador em questão é o MLP o valor obtido inicialmente é de 94,9. Com o seletor Relief o mesmo se mantêm, com o PCA a acurácia atinge 71,4 e com o CP o índice passa a ser de 98,8. Para os dados obtidos pelo RBF tem se que a acurácia obtida sem seleção de características foi de 82,7, com a utilização do Relief foi de 83,2, com o PCA 74,8 e com o CP foi de 98,6. Com o SVM os valores obtidos foram de 97,6 - sem nenhuma seleção de características, 97,6 com o Relief, 74,8 com o PCA e 98,7 com o CP. Para a média destes cinco classificadores as acurácias obtidas foram de 90,2 sem utilização de seletores, 90,3 com o Relief, 71,2 com o PCA e 95,3 com o CP. Os valores da acurácia média estão apresentados na tabela 6.9 , a variância da acurácia está apresentada na tabela 6.10 e representados na figura 6.9.

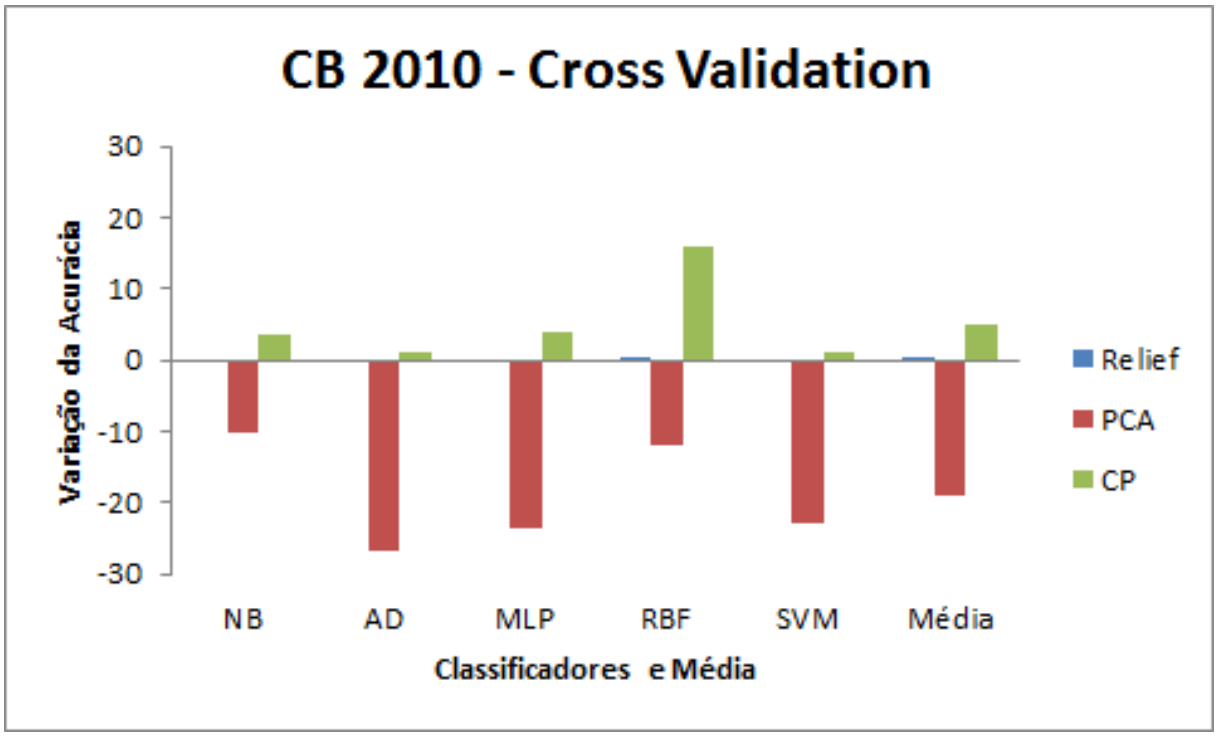


Figura 6.9: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010.

Tabela 6.9: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010.

	NB	AD	MLP	RBF	SVM	Média
Base	79%	97,1%	94,9%	82,7%	97,6%	90,2%
Relief	79%	97,1%	94,9%	83,2%	97,6%	90,3%
PCA	68,8%	70,5%	71,4%	70,9%	74,8%	71,2%
CP	82,6%	98,2%	98,8%	98,6%	98,7%	95,3%

Tabela 6.10: Variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	0	0	0,5	0	0,1
PCA	-10,2	-26,6	-23,5	-11,8	-22,8	-19
CP	3,6	1,1	3,9	15,9	1,1	5,1

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final 5,1. O Relief manteve o desempenho em 4 classificadores e obteve ganho em 1, obtendo 0,1 como média. O PCA piorou os resultados dos classificadores nos 5 casos, obtendo como variação da acurácia -19.

No campeonato brasileiro de 2012 considerando o classificador NB tem-se acurácia obtida de 82,1 sem o uso de seleção de características, com o método de seleção Relief houve uma perda de 0,1, com o PCA uma perda de 12,2 e com o CP um aumento de 0,9. Com a AD a acurácia original foi de 97,0, com o Relief a mesma se manteve, com o PCA houve uma queda de 28,1 e por sua vez com o CP houve melhora de 0,2. No caso do MLP a acurácia foi de 94,9 sem a utilização de métodos de seleção. Com o Relief a mesma permaneceu inalterada, com o PCA houve perda de 26,3 e com o CP houve melhoria de 3,3. Para o RBF o índice obtido foi de 84,3 sem o uso de ferramentas de seleção de características, de 85,4 com o Relief, 71,7 com o PCA e 98,6 com o CP. Já para o SVM o valor original obtido foi de: 97,3, com o Relief de 97,2, de 74,6 com o PCA e de 98,3 com o uso do CP. Finalmente as médias obtidas pelos cinco classificadores tem se que a acurácia obtida com todas as características foi de: 91,1, com o Relief 91,3, com o PCA 70,7 e com o CP 94,9. Os valores da acurácia média estão apresentados na tabela 6.11 , a variância da acurácia está apresentada na tabela 6.12 e representados na figura 6.10.

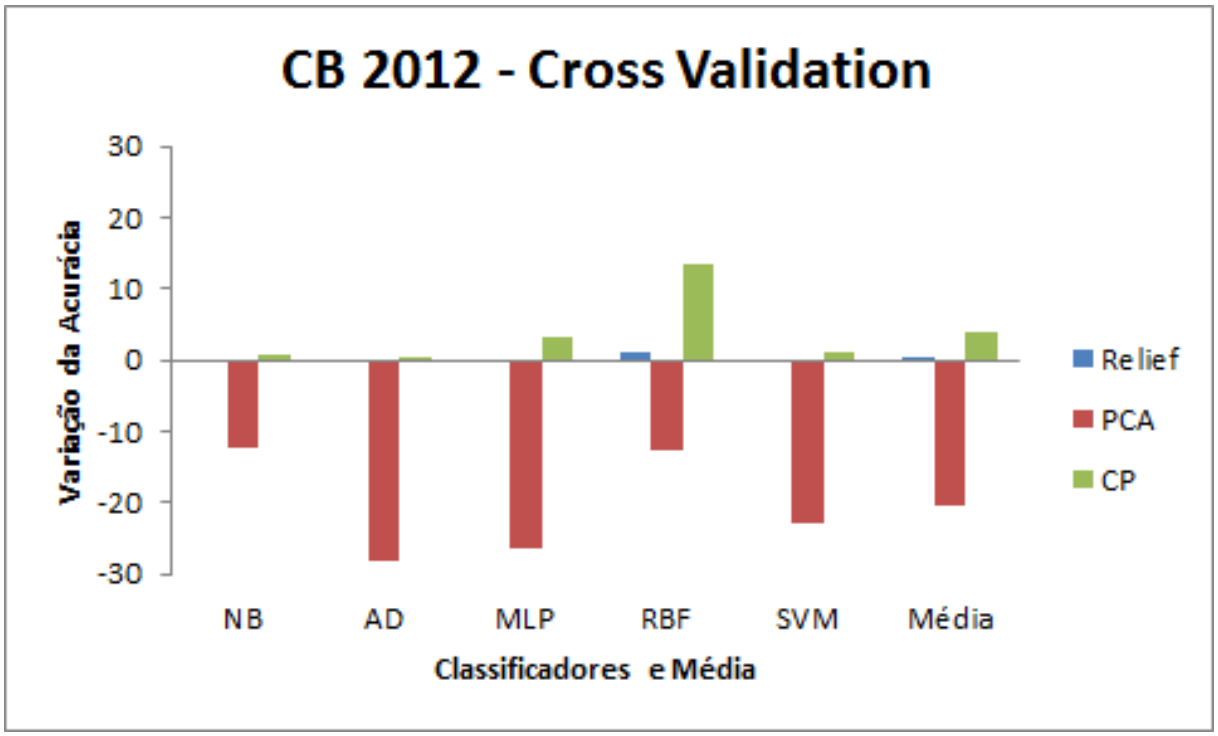


Figura 6.10: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012.

Tabela 6.11: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012.

	NB	AD	MLP	RBF	SVM	Média
Base	82,1%	97%	95,2%	84,3%	97,3%	91,1%
Relief	82%	97%	95,2%	85,4%	97,2%	91,3%
PCA	69,9%	68,9%	68,9%	71,7%	74,6%	70,7%
CP	83%	97,2%	98,5%	97,6%	98,3%	94,9%

Tabela 6.12: Variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012.

	NB	AD	MLP	RBF	SVM	Média
Relief	-0,1	0	0	1,1	-0,1	0,2
PCA	-12,2	-28,1	-26,3	-12,6	-22,7	-20,4
CP	0,9	0,2	3,3	13,3	1	3,8

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final 3,8. O Relief manteve o desempenho em 2 classificadores e obteve ganho em 2 e perdeu em 1 dos casos, obtendo 0,2 como média. O PCA piorou os resultados dos classificadores nos 5 casos, obtendo como variação da acurácia -20,4.

No campeonato inglês temporada de 2014/15 foram obtidos os seguintes resultados: no classificador NB a acurácia obtida foi de 80,7 sem a utilização de seleção de características, não havendo alteração quando se utiliza os seletores Relief e PCA e uma melhora de 0,4 quando se utiliza o CP; no classificador AD sem a utilização de seletores a acurácia obtida foi de 99,4, esse índice permanece o mesmo tanto com o uso do Relief quanto com o uso do PCAe uma melhora de 0,2 quando se utiliza o CP; para o classificador MLP a acurácia inicial obtida foi de 100,0 e se manteve com os três seletores de características; já para o RBF obteve se uma acurácia de 97,4 sem a utilização de seletores. A mesma se manteve com a utilização do Relief, uma melhora de 0,4 com o uso do PCA e de 2,3 com o CP; no SVM a acurácia inicial foi de 98,6, com o aumento de performance de 0,1 com o Relief e PCA e de 1,4 com o CP. Quando se leva em conta os valores médios obtidos por esses cinco classificadores tem se que a acurácia inicialmente obtida foi de 95,2, tendo esse índice se mantido com o Relief, um aumento de 0,1 com o PCA e de 0,8 com o CP. Os valores da acurácia média estão apresentados na tabela 6.13 , a variância da acurácia está apresentada na tabela 6.14 e representados na figura 6.11.

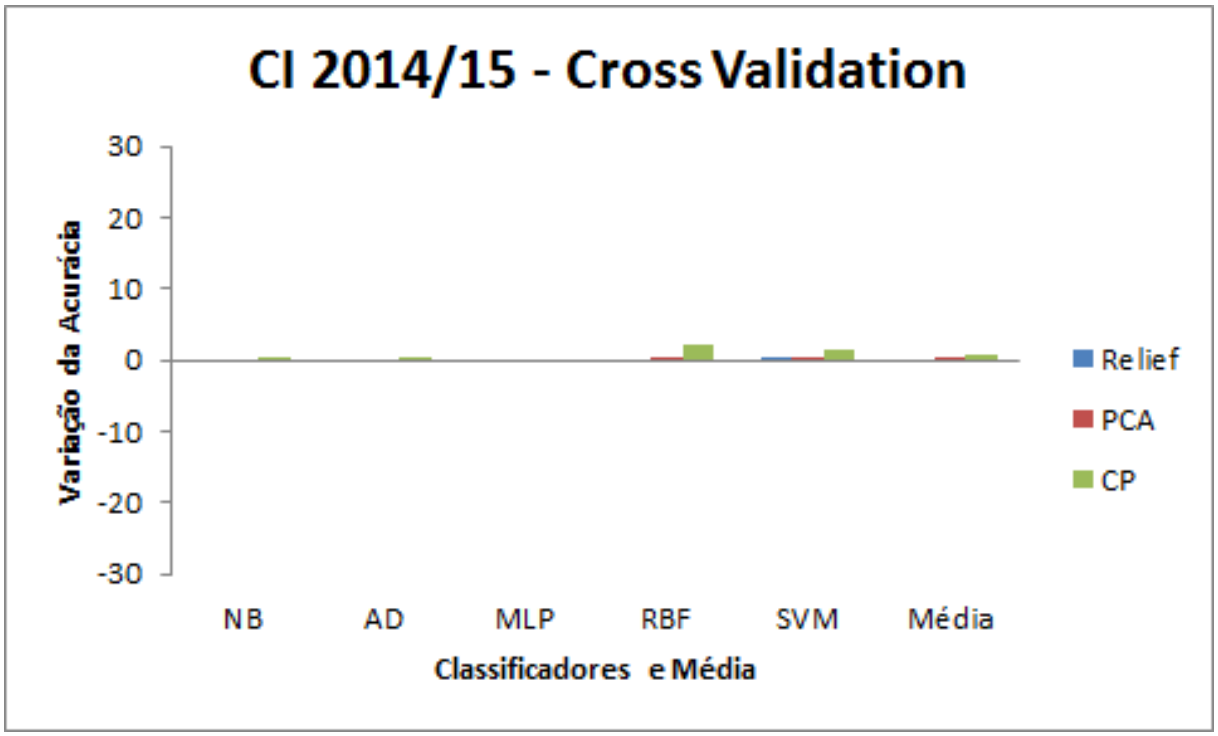


Figura 6.11: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Inglês temporada de 2014/15.

Tabela 6.13: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Inglês temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Base	80,7%	99,4%	100%	97,4%	98,6%	95.2%
Relief	80,7%	99,4%	100%	97,4%	98,7%	95,2%
PCA	80,7%	99,4%	100%	97,8%	98,7%	95,3%
CP	81,1%	99,6%	100%	99,7%	100%	96%

Tabela 6.14: Variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Inglês temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	0	0	0	0,1	0,02
PCA	0	0	0	0,4	0,1	0,1
CP	0,4	0,2	0	2,3	1,4	0,8

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final 0,8. O Relief manteve o desempenho em 4 classificadores e obteve ganho em 1 dos casos, obtendo 0,02 como média. O PCA manteve o desempenho em 3 classificadores e obteve ganho em 2 dos casos obtendo como variação da acurácia 0,1.

No caso do campeonato espanhol temporada de 2014/15 foram obtidos os seguintes resultados: utilizando se o classificador NB sem o uso de seletores de características a acurácia obtida foi de 83,1, não houvendo alteração com a utilização do Relief, uma queda de 0,6 com o uso do PCA e uma melhora de 3,9 com o CP; para a AD e o MLP as acurácias obtidas inicialmente foram de 99,4 e 100,0 e não houve variação com o uso dos seletores de características; para o RBF a acurácia inicial foi de 97,3 e um aumento de 0,3 para o Relief, 0,8 para o PCA e 2,1 para o CP; o RBF obteve 93,4 inicialmente, valor esse mantido com o Relief, uma queda de 0,7 com o PCA e um aumento de 4,9 com o CP. A média dos cinco classificadores foi de 94,6, com uma melhora de 0,1 com o Relief, uma queda de 0,1 com o PCA e uma melhora de 2,2 com o CP. Os valores da acurácia média estão apresentados na tabela 6.15 , a variância da acurácia está apresentada na tabela 6.16 e representados na figura 6.12.

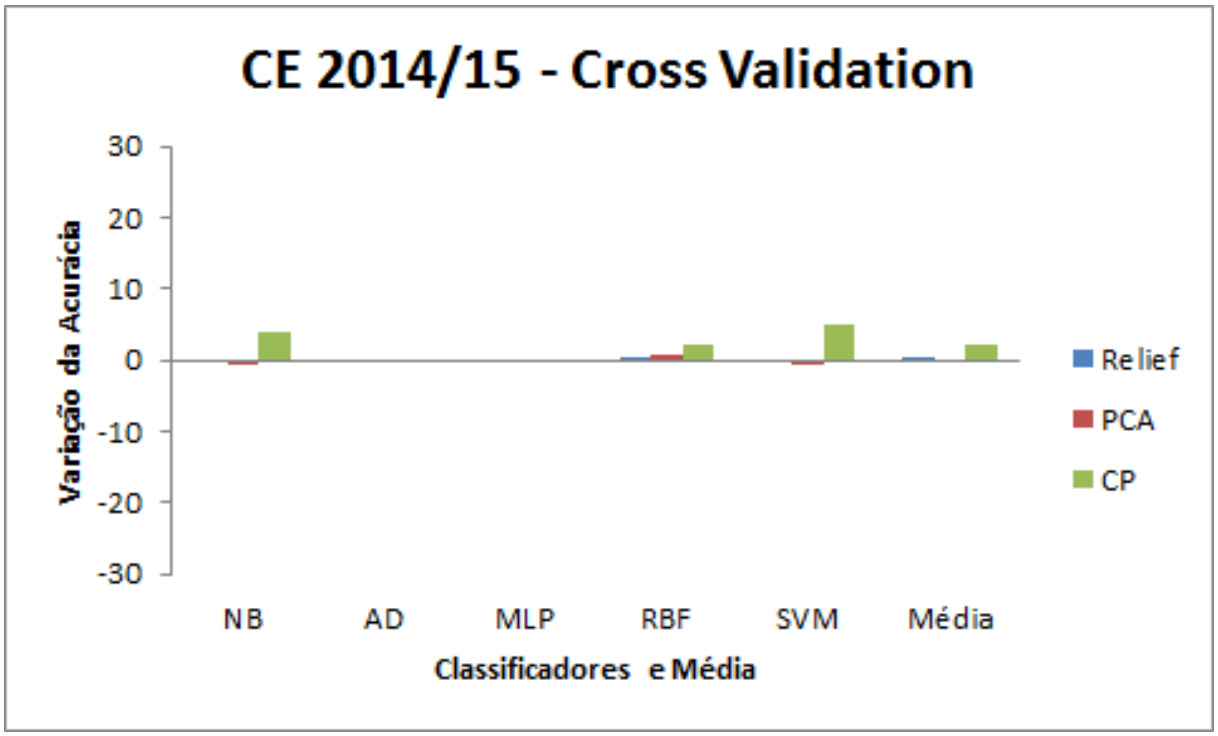


Figura 6.12: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

Tabela 6.15: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Base	83,1%	99,4%	100%	97,3%	93,4%	94,6%
Relief	83,1%	99,4%	100%	97,6%	93,4%	94,7%
PCA	82,5%	99,4%	100%	98,1%	92,7%	94,5%
CP	87%	99,4%	100%	99,4%	98,3%	96,8%

Tabela 6.16: Variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	0	0	0,3	0	0,1
PCA	-0,6	0	0	0,8	-0,7	-0,1
CP	3,9	0	0	2,1	4,9	2,2

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia de três classificadores envolvidos e manteve em 2 obtendo como variação da acurácia final 2,2. O Relief manteve o desempenho em 4 classificadores e obteve ganho em 1 dos casos, obtendo 0,1 como média. O PCA piorou o desempenho em 2 classificadores, obteve ganho em 1 dos casos e manteve nos outros 2 obtendo como variação da acurácia -0,1. É possível afirmar que para os 20 casos analisados para a técnica *cross validation* melhorou o desempenho em 18 e manteve em 2 e para as quatros bases de dados foi o que obteve o melhor ganho na variação da acurácia. O Relief melhorou em 5 casos, piorou em 2 e manteve para 14 dos casos, sendo segunda melhor variação das acurácias. O PCA melhorou 4 casos, piorou 11 e melhorou 5, sendo o pior desempenho em relação a variação das acurácias.

6.3.2 Resultados com o Slinding Window

A técnica slinding window foi aplicada novamente nas 4 base de dados anteriores: Campeonato Brasileiro de 2010 e 2012, Campeonato Inglês 2014/15 e Campeonato Espanhol 2014/15. Foi obtida uma média considerando as 3 combinações possíveis: vitória x empate, vitória x derrota e empate x derrota. As médias obtidas foram comparadas para que seja possível observar a variação obtida em cada um dos casos.

O Campeonato Brasileiro de 2010 apresentou os seguintes resultados: com a utilização do classificador NB e sem o uso de seletores de características a acurácia obtida foi de 67,0, com o seletor Relief a acurácia se manteve a mesma, com o PCA houve piora no desempenho de 7,4 e com o CP houve melhora de 4,7; na AD a acurácia inicial encontrada foi de 75,7.

Com o seletor Relief houve piora de 0,2, com o PCA a queda foi de 16 e com o CP houve melhora de 15,5; para o MLP a acurácia sem seleção de características foi de 74,0, um ganho de 0,9 com o uso do Relief, uma queda de 16,4 com o PCA e uma melhora de 23,8 com o CP; com o RBF o índice inicial foi de 68,2, perdas de 0,4 com o Relief e de 5,5 com o PCA e um ganho de 27,2 com o CP; já no SVM a acurácia inicial foi de 73,6, o desempenho se manteve com o Relief, uma piora de 9,7 com o PCA e um ganho de 15,6 com o CP. Considerando a média obtida pelos cinco classificadores obteve se uma acurácia de 71,7. Com o Relief esse valor se manteve, no caso do PCA houve queda de rendimento de 11,0 e com o CP houve melhora de desempenho de 17,2. Os valores da acurácia média estão apresentados na tabela 6.17 , a variância da acurácia está apresentada na tabela 6.18 e representados na figura 6.13.

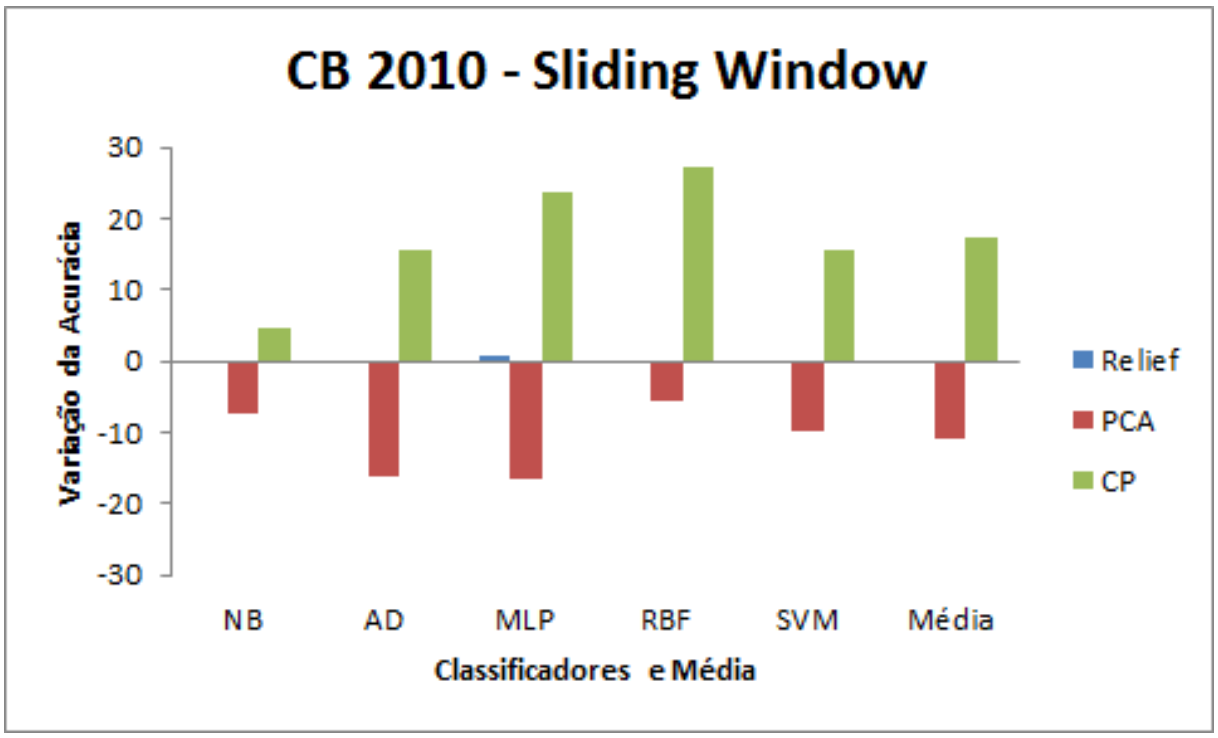


Figura 6.13: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

Tabela 6.17: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

	NB	AD	MLP	RBF	SVM	Média
Base	67%	75,7%	74%	68,2%	73,6%	71,7%
Relief	67%	75,5%	74,9%	67,8%	73,6%	71,7%
PCA	59,6%	59,7%	57,6%	62,7%	63,9%	60,7%
CP	71,7%	91,2%	97,8%	95,4%	89,2%	88,9%

Tabela 6.18: Variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	-0,5	0	4,4	0,1	0,2
PCA	-5,4	-20,2	-13,4	0,5	-10	-10,2
CP	5,6	16,1	25,4	34,1	16,4	19

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final de 19. O Relief manteve o desempenho em 2 classificadores, obteve ganho em 2 dos casos e manteve em 1 obtendo 0,2 como média. O PCA piorou 4 dos casos e melhorou 1 deles obtendo como variação da acurácia -10,2.

No caso do Campeonato Brasileiro de 2012 os resultados obtidos foram: com o NB sem o uso de seleção de características a acurácia obtida foi de 66,5, com a utilização do Relief não houve alteração, com o PCA houve uma piora no desempenho de 5,4 e uma melhora com o CP de 5,6; com o AD a acurácia foi de 76,6 sem o uso de seletores, houve uma queda com o uso do Relief de 0,5, com o PCA a perda de desempenho foi de 20,2, e com o CP houve melhora de 16,1; no caso do MLP o índice original foi de 71,8, valor esse mantido com o Relief, houve queda de 13,4 com o PCA e aumento de 25,4 com o CP; para o RBF a acurácia inicial foi de 61,1, com a utilização do Relief houve melhora de 4,4, com o PCA a

melhora foi de 0,5 e com o CP foi de 34,1; já para o caso do SVM a acurácia inicial foi de 71,9, com melhora de 0,1 para o Relief, queda de 10,0 para o PCA e aumento de 16,4 para o CP. Considerando as médias obtidas pelos cinco classificadores tem se que a acurácia base foi de 70,1, com aumento de 0,2 com o Relief, queda de 10,2 com o PCA e aumento de 19 com o CP. Os valores da acurácia média estão apresentados na tabela 6.19 , a variância da acurácia está apresentada na tabela 6.20 e representados na figura 6.14.

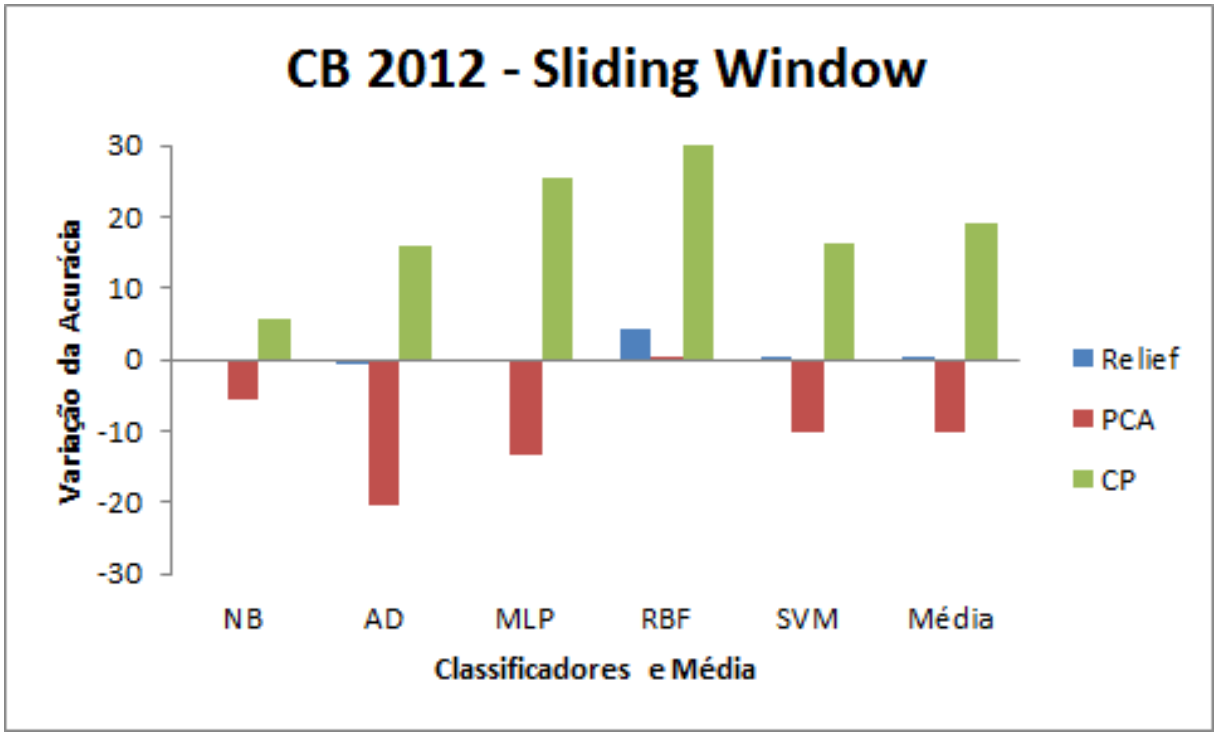


Figura 6.14: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012.

Tabela 6.19: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012.

	NB	AD	MLP	RBF	SVM	Média
Base	66,5%	76,6%	71,8%	61,1%	71,9%	70,1%
Relief	66,5%	76,1%	71,8%	65,5%	72%	70,3%
PCA	61,1%	56,4%	58,4%	61,6%	61,9%	59,9%
CP	72,1%	92,7%	97,2%	95,2%	88,3%	89,1%

Tabela 6.20: Variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	-0,2	0,9	-0,4	0	0
PCA	-7,4	-16	-16,4	-5,5	-9,7	-11
CP	4,7	15,5	23,8	27,2	15,6	17,2

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final de 17,2. O Relief manteve o desempenho em 2 classificadores, obteve ganho em 1 dos casos e perdeu em 2 obtendo 0 como média. O PCA piorou os 5 obtendo como variação da acurácia -11.

Para a base de dados do Campeonato Inglês temporada de 2014/15 os resultados obtidos foram: com o classificador NB a acurácia obtida foi de 70,3 sem a presença de seletores, valor mantido com a presença do Relief, com ganho de performance de 0,8 com o PCA e de 0,7 com o CP; no caso da AD a acurácia inicial foi de 85,6, valor mais uma vez mantido com o Relief, com melhora de 2,1 para o PCA e de 4,5 para o CP; com o MLP o índice inicial foi de 87,8, valor mantido com o Relief, melhorado em 5,2 com o PCA e 11,3 com o CP; para o RBF a acurácia inicial foi de 78,1, com queda de 0,2, aumento de 2,9 com o PCA e 15,8 com o CP; no caso do SVM a acurácia base foi de 79,3, com aumento de 0,4 com o Relief, de 2,0 com o PCA e 5,4 com o CP. Considerando a média dos cinco

classificadores obteve se 80,2, valor mantido com o Relief, com aumento de 2,6 com o PCA e 7,5 com o CP. Os valores da acurácia média estão apresentados na tabela 6.21 , a variância da acurácia está apresentada na tabela 6.22 e representados na figura 6.15.

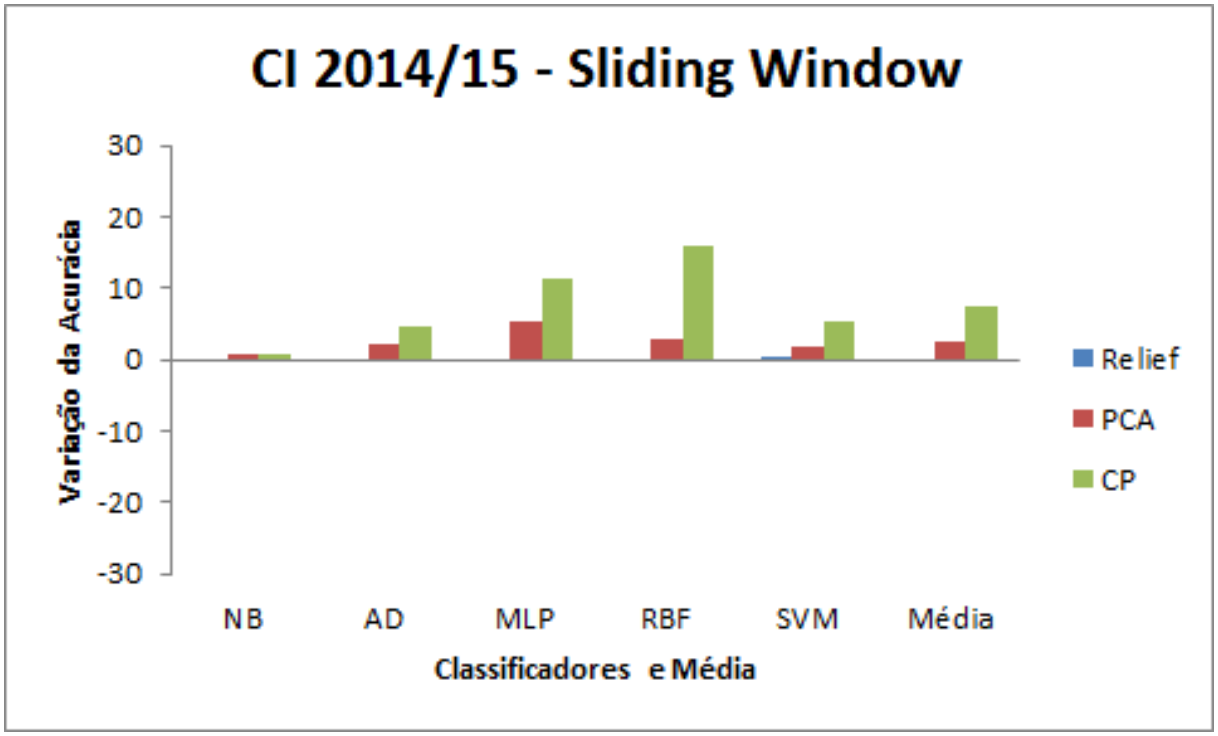


Figura 6.15: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Inglês temporada de 2014/15.

Tabela 6.21: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Inglês temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Base	70,3%	85,6%	87,8%	78,1%	79,3%	80,2%
Relief	70,3%	85,6%	87,8%	77,9%	79,7%	80,2%
PCA	71,1%	87,7%	93%	81%	81,3%	82,8%
CP	71 %	90,1%	99,1%	93,9%	84,7%	87,7%

Tabela 6.22: Variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Inglês temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	0	0	-0,2	0,4	0
PCA	0,8	2,1	5,2	2,9	2	2,6
CP	0,7	4,5	11,3	15,8	5,4	7,5

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia dos cinco classificadores envolvidos e obteve como variação da acurácia final de 7,5. O Relief manteve o desempenho em 3 classificadores, obteve ganho em 1 dos casos e perdeu em 1 obtendo 0 como média. O PCA melhorou os 5 casos obtendo como variação da acurácia 2,6.

Para a base de dados do Campeonato Espanhol temporada de 2014/15 os resultados alcançados foram: para o NB sem a utilização de seletores de características a acurácia obtida foi de 75,1, mesmo valor alcançado com o Relief e também com o PCA, com o CP houve piora de 0,2; com a AD acurácia inicial foi de 86,1, melhoria de 0,2 com o uso do Relief, melhora de 0,1 com o PCA e de 4,8 com o CP; para o MLP obteve se 84,6, com aumento de 0,1 com o Relief, de 4,8 com o PCA e 11,4 com o CP. Com o RBF a acurácia inicial foi de 79,3, com ganho de 1,3 para o Relief, 2,6 para o PCA e 14,2 com o CP; com o SVM o índice original foi de 78,6 com aumento de 0,2 com o uso do Relief, 0,6 com o PCA e 4,4. com o CP. A média de acurácia dos cinco classificadores sem seleção de características foi de 80,7, com ganho de 0,4 com o Relief, 1,6 com o PCA e 6,9 com o CP. Os valores da acurácia média estão apresentados na tabela 6.23 , a variância da acurácia está apresentada na tabela 6.24 e representados na figura 6.16.

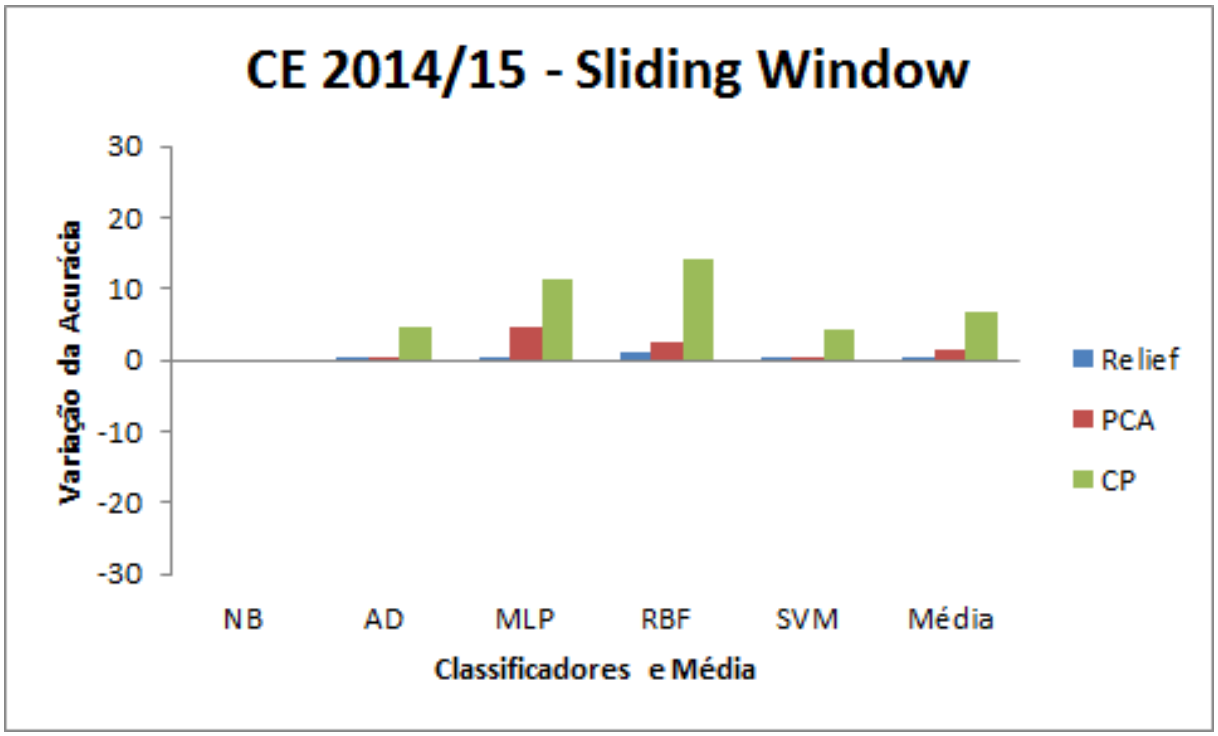


Figura 6.16: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

Tabela 6.23: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Base	75,1%	86,1%	84,6%	79,3%	78,6%	80,7%
Relief	75,1%	86,3%	84,7%	80,6%	78,8%	81,1%
PCA	75,1%	86,2%	89,4%	81,9%	79,2%	82,3%
CP	74,9%	90,9%	96%	93,5%	83%	87,6%

Tabela 6.24: Variação das acurácias obtidas pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Espanhol temporada de 2014/15.

	NB	AD	MLP	RBF	SVM	Média
Relief	0	0,2	0,1	1,3	0,2	0,4
PCA	0	0,1	4,8	2,6	0,6	1,6
CP	-0,2	4,8	11,4	14,2	4,4	6,9

Analisando os resultados obtidos é possível afirmar que o CP melhorou a acurácia de 4 classificadores envolvidos e perdeu em 1 obtendo como variação da acurácia final 6,9. O Relief melhorou o desempenho em 4 classificadores e manteve em 1 dos casos, obtendo 0,4 como média. O PCA manteve o desempenho em 1 classificador e obteve ganho em 4 dos casos obtendo como variação da acurácia 1,6. É possível afirmar que para os 20 casos analisados para a técnica *cross validation* melhorou o desempenho em 19 e piorou em 1 e para as quatros bases de dados foi o que obteve o melhor ganho na variação da acurácia. O Relief melhorou em 8 casos, piorou em 4 e manteve para 8 dos casos, sendo segunda melhor variação das acurácias. O PCA melhorou 11 casos, empatou em 3 e piorou em 6, sendo o pior desempenho em relação a variação das acurácias.

6.4 Resultados obtidos para os Campeonatos Brasileiro de 2010 e 2012 com as características reduzidas

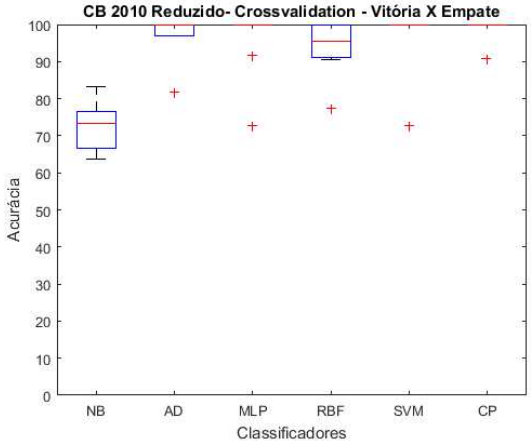
Foram realizados testes com as bases de dados do CB 2010 e CB 2012 com as caracteríticas reduzidas ao vetor de características dos CI 2014/15 e CE 2014/15, ou seja ao invés das 54 características originais, os testes foram realizados com 18 características. Os resultados obtidos para a técnica *cross validation* se encontram na subseção 6.4.1 e na subseção 6.4.2 aparecem os resultados para a técnica *sliding window*.

6.4.1 Cross Validation

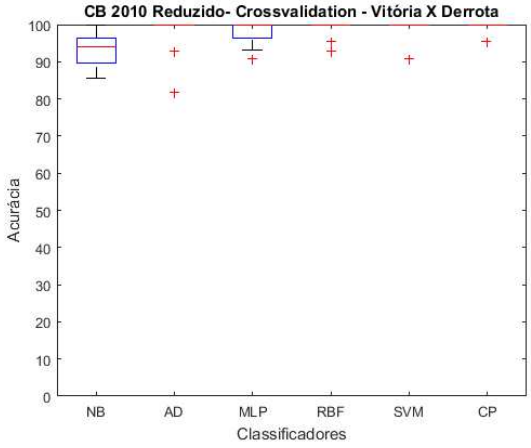
A base do campeonato brasileiro 2010 com as características reduzidas apresentou as seguintes acurácias: CP(99,09±2,87, 99,54±1,43 e 99,37±1,97) , SVM(97,27±8,6, 99,09±2,8 e 98,75±3,9), RBF(94,24±7,0, 98,83±2,5 e 91,80±5,0), MLP(96,48±8,7, 98,04±3,4 e 98,27±4,0), AD(97,58±5,6, 97,46±5,9 e 97,25±3,8) e NB(72,42±6,5, 93,81±4,6 e 75,78±8,1) para as combinações vitória x empate, vitória x derrota e empate x derrota. Tais valores são apresentados na tabela 6.25. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.17.

Tabela 6.25: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010 com as características reduzidas.

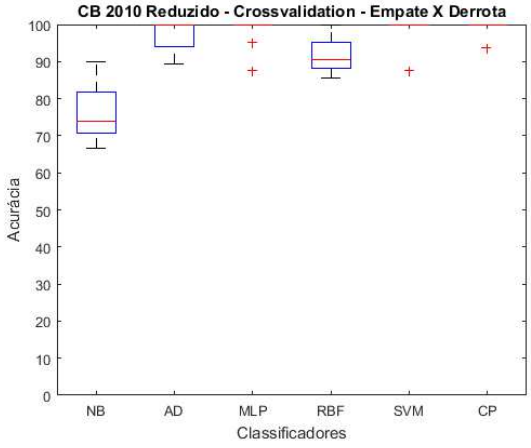
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	72,42±6,5	93,81±4,6	75,78±8,1
AD	97,58±5,6	97,46±5,9	97,25±3,8
MLP	96,48±8,7	98,04±3,4	98,27±4,0
RBF	94,24±7,0	98,83±2,5	91,80±5,0
SVM	97,27±8,6	99,09±2,8	98,75±3,9
CP	99,09±2,87	99,54±1,43	99,37±1,97



(a)



(b)



(c)

Figura 6.17: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010 com as características reduzidas.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.25, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(99,09±2,87), AD(97,58±5,6), SVM(97,27±8,6), MLP(96,48±8,7), RBF(94,24±7,0) e NB(72,42±6,5). Através do gráfico 6.17a podemos observar que CP, AD, SVM e MLP possuem a mediana ideal com a diferença de desempenho sendo explicada pelos valores discrepantes. Com RBF e NB com desempenhos inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(99,54±1,43), SVM(99,09±2,8), RBF(98,83±2,5), MLP(98,04±3,4), AD(97,46±5,9) e NB(93,81±4,6). Através do gráfico 6.17b temos que CP, SVM, MLP, AD e RBF atingiram a mediana ideal com a diferença de desempenho sendo novamente explicada pelos valores discrepantes. No caso do NB houve melhora em relação a combinação anterior mas o desempenho segue inferior aos demais classificadores. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(99,37±1,97), SVM(98,75±3,9), MLP(98,27±4,0), AD(97,25±3,8), RBF(91,80±5,0) e NB(75,78±8,1). Através do gráfico 6.17c temos que CP, SVM, MLP e AD atingiram a mediana ideal com a diferença de desempenho sendo novamente explicada pelos valores discrepantes. Nos casos do MLP e NB houve desempenho similar a combinação vitória x empate.

Em relação a comparação entre o conjunto original de 54 características para a redução de 18 a análise será realizada de três maneiras distintas: uma tabela 6.27 contendo a média da acurácia das três combinações (vitória x empate, vitória x derrota, empate x derrota), uma tabela 6.26 contendo a varição da acurácia , diferença entre o valor base e o valor obtido por cada um dos três seletores de características(CP, PCA e Relief), e uma figura 6.18 contendo um gráfico de barras com variação da acurácia distribuída pelos classificadores.

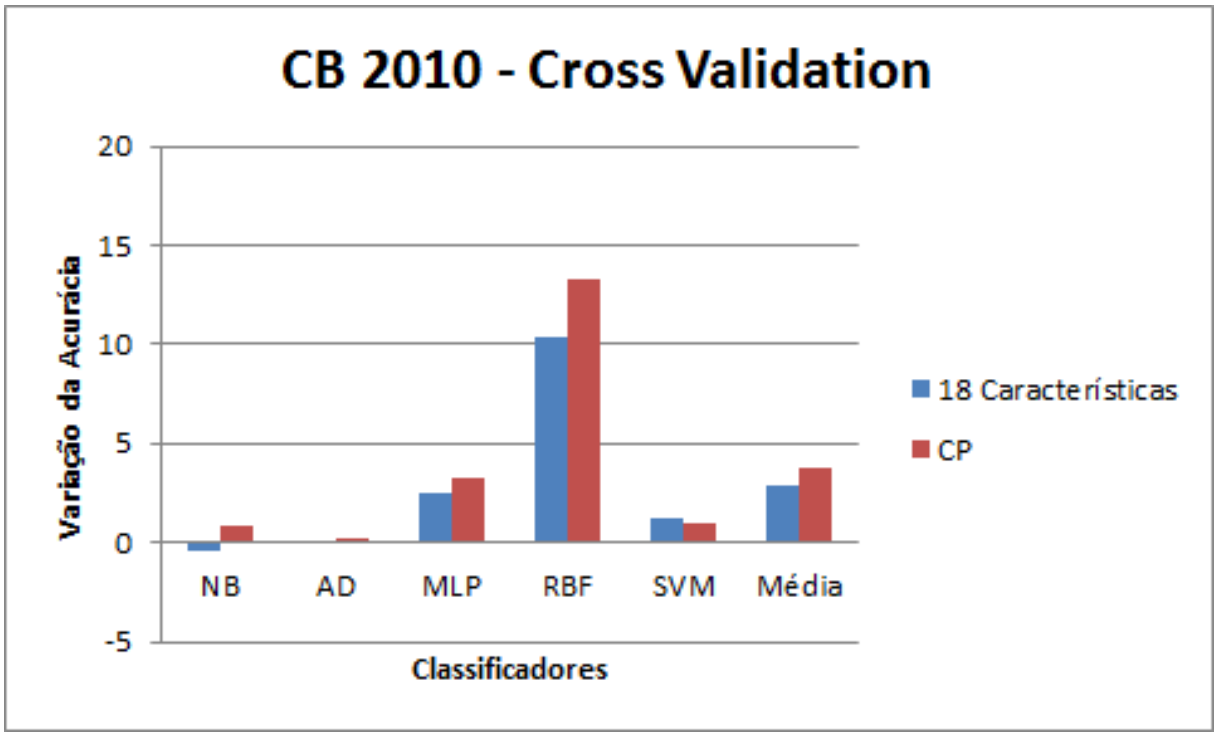


Figura 6.18: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2010 com as características reduzidas.

Tabela 6.26: Resultados obtidos pelos classificadores utilizando a redução para 18 características para o CB 2010 com a técnica cross validation.

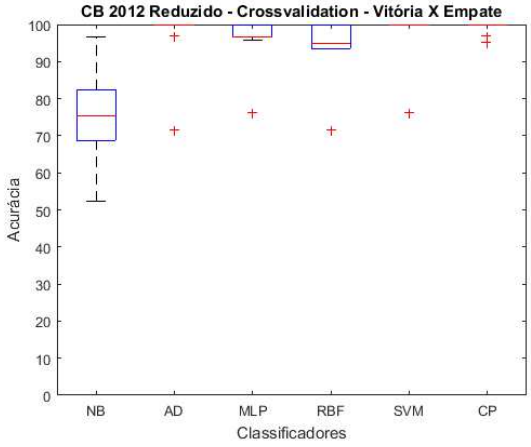
	NB	AD	MLP	RBF	SVM	Média
54 Características	82,1	97	95,2	84,3	97,3	91,1
CP	83	97,2	98,5	97,6	98,3	94,9
18 Características	81,67	97,03	97,72	94,65	98,52	93,92

Tabela 6.27: Variação das acurácias obtidas pelos classificadores utilizando a redução para 18 características para o CB 2010 com a técnica cross validation.

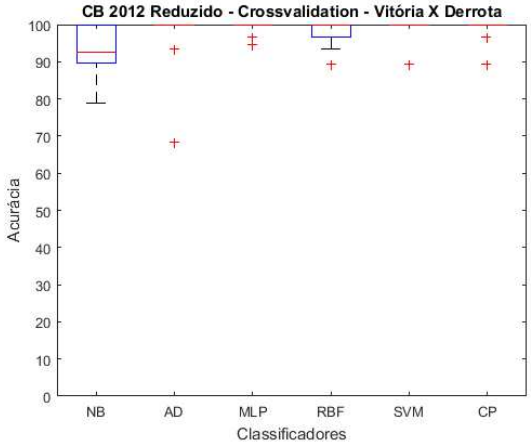
	NB	AD	MLP	RBF	SVM	Média
18 Características	-0,43	0,03	2,52	10,35	1,22	2,82
CP	0,9	0,2	3,3	13,3	1	3,8

No campeonato CB 2010, com a técnica *cross validation* foram obtidos os seguintes resultados: no classificador NB a acurácia obtida foi de 82,1 com as 54 características originais, 83 para as 18 características, piorado em 0,43 e foi de 83 com o CP, aumentando em 0,9. Com a AD a acurácia obtida foi de 97 com as 54 características originais, 97,03 para as 18 características, aumentando em 0,03 e foi de 97,2 com o CP, aumentando em 0,2. Para o MLP a acurácia obtida foi de 95,2 com as 54 características originais, 97,72 para as 18 características, aumentando em 2,52 e foi de 98,5 com o CP, aumentando em 3,3. Com o RBF a acurácia obtida foi de 84,3 com as 54 características originais, 94,65 para as 18 características, aumentando em 10,38 e foi de 97,6 com o CP, aumentando em 13,3. Para o SVM a acurácia obtida foi de 97,3 com as 54 características originais, 98,53 para as 18 características, aumentando em 1,22 e foi de 98,3 com o CP, aumentando em 1. O ganho na acurácia média para as 18 características foi de 2,82 e de 3,8 para o CP. Os valores da acurácia média estão apresentados na tabela 6.26 , a variância da acurácia está apresentada na tabela 6.27 e representados na figura 6.18.

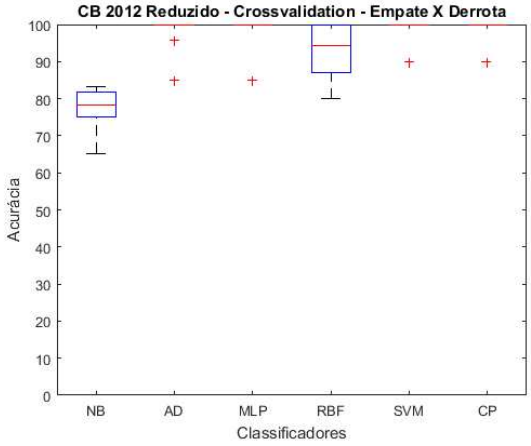
A base do campeonato brasileiro 2012 com as características reduzidas apresentou as seguintes acurácias: CP(99,21±1,7, 98,62±3,3 e 99,00±3,1) , SVM(97,61±7,5, 98,94±3,3 e 99,00±3,1), RBF(93,83±8,3, 97,95±3,6 e 92,15±7,3), MLP(95,52±6,9, 99,12±1,8 e 98,50±4,7), AD(96,83±8,9, 96,19±9,9 e 98,06±4,7) e NB(75,15±12,0, 92,83±6,7 e 77,01±5,7) para as combinações vitória x empate, vitória x derrota e empate x derrota. Tais valores são apresentados na tabela 6.28. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.19.



(a)



(b)



(c)

Figura 6.19: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

Tabela 6.28: Resultados obtidos pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	75,15±12,0	92,83±6,7	77,01±5,7
AD	96,83±8,9	96,19±9,9	98,06±4,7
MLP	95,52±6,9	99,12±1,8	98,50±4,7
RBF	93,83±8,3	97,95±3,6	92,15±7,3
SVM	97,61±7,5	98,94±3,3	99,00±3,1
CP	99,21±1,7	98,62±3,3	99,00±3,1

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.28, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(99,21±1,7), SVM(97,61±7,5), AD(96,83±8,9), MLP(95,52±6,9), RBF(93,83±8,3) e NB(75,15±12,0). Através do gráfico 6.19a temos que CP, SVM, e AD atingiram a mediana ideal com a diferença de desempenho explicada pelos valores discrepantes. Nos casos do MLP e RBF houve desempenho similar com vantagem para o primeiro em relação a mediana e aos quartis e limites inferiores. O desempenho de NB destoa dos demais. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: MLP(99,12±1,8), SVM(98,94±3,3), CP(98,62±3,3), RBF(97,95±3,6), AD(96,19±9,9) e NB(92,83±6,7). Através do gráfico 6.19b temos que MLP, SVM, CP, RBF e AD com a mediana ideal e com a diferença de desempenho sendo explicada pelos valores discrepantes e com NB tendo tido uma melhora de desempenho em relação a combinação anterior mas ainda abaixo dos demais. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(99,00±3,1), SVM(99,00±3,1), MLP(98,50±4,7), AD(98,06±4,7), RBF(92,15±7,3) e NB(77,01±5,7). Através do gráfico 6.19c temos que CP, SVM, MLP e AD com valores concentrados em torno da mediana ideal, com a diferença de desempenho sendo explicada pelos valores discrepantes. Com performance um pouco mais abaixo podemos observar RBF e NB.

Em relação a comparação entre o conjunto original de 54 características para a redução de 18 a análise será realizada de três maneiras distintas: uma tabela 6.32 contendo a média da acurácia das três combinações (vitória x empate, vitória x derrota, empate x derrota), uma tabela contendo a varição da acurácia 6.33, diferença entre o valor base e o valor obtido por cada um dos três seletores de características(CP, PCA e Relief), e uma figura 6.20 contendo um gráfico de barras com variação da acurácia distribuída pelos classificadores.

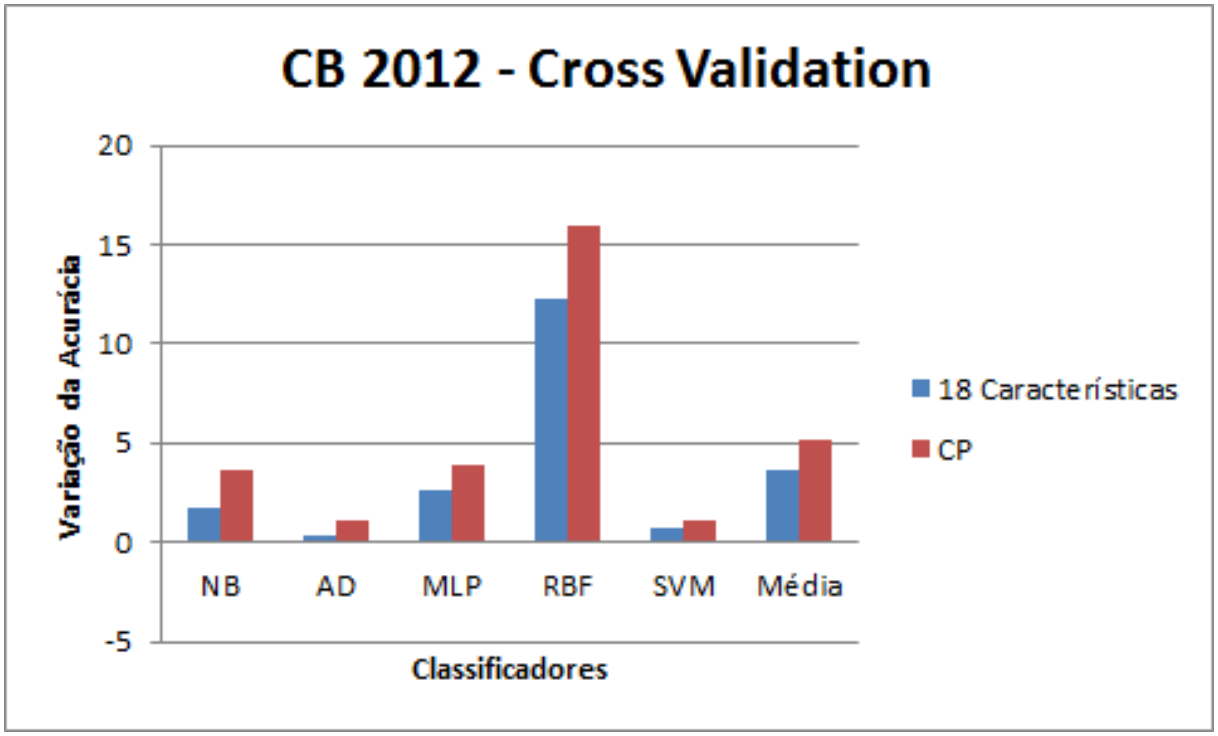


Figura 6.20: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

Tabela 6.29: Resultados obtidos pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica cross validation.

	NB	AD	MLP	RBF	SVM	Média
54 Características	79	97,1	94,9	82,7	97,6	90,2
CP	82,6	98,2	98,8	98,6	98,7	95,3
18 Características	80,68	97,44	97,59	94,96	98,37	93.81

Tabela 6.30: Variação das acurácias obtidas pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica cross validation.

	NB	AD	MLP	RBF	SVM	Média
18 Características	1,68	0,34	2,69	12,26	0,77	3,61
CP	3,6	1,1	3,9	15.9	1,1	5,1

No campeonato CB 2012 foram obtidos os seguintes resultados: no classificador NB a acurácia obtida foi de 79 com as 54 características originais, 80,68 para as 18 características, aumentando em 1,68 e foi de 82,6 com o CP, aumentando em 3,6. Com a AD a acurácia obtida foi de 97,1 com as 54 características originais, 97,44 para as 18 características, aumentando em 0,34 e foi de 98,2 com o CP, aumentando em 1,1. Para o MLP a acurácia obtida foi de 94,9 com as 54 características originais, 97,59 para as 18 características, aumentando em 12,26 e foi de 98,6 com o CP, aumentando em 3,9. Com o RBF a acurácia obtida foi de 82,7 com as 54 características originais, 94,96 para as 18 características, aumentando em 12,26 e foi de 98,6 com o CP, aumentando em 15,9. Para o SVM a acurácia obtida foi de 97,6 com as 54 características originais, 98,37 para as 18 características, aumentando em 3,61 e foi de 98,7 com o CP, aumentando em 5,1. O ganho na acurácia média para as 18 características foi de 3,61 e de 5,1 para o CP. Os valores da acurácia média estão apresentados na tabela 6.29 , a variância da acurácia está apresentada na tabela 6.30 e representados na figura 6.20. Analisando os resultados obtidos pode se afirmar que para essa base de dados na técnica *cross validation* a redução para 18 características melhorou o desempenho para todos os caso

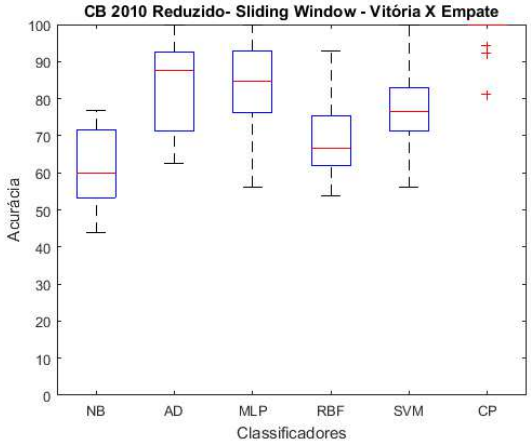
mas não superou o desempenho do CP.

6.4.2 Slinding Window

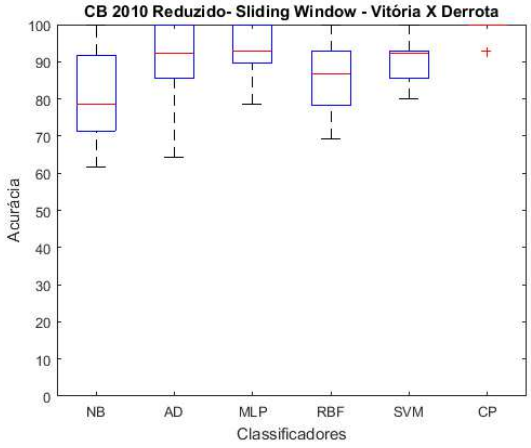
A base do campeonato brasileiro 2010 com as características reduzidas apresentou as seguintes acurácias: CP(98,11±4,8, 99,15±2,3 e 98,46±3,4) , SVM(77,97±11,3, 91,39±5,6 e 73,48±11,5), RBF(69,29±10,9, 85,85±9,6 e 64,89±14,7), MLP(83,45±12,4, 92,95±6,6 e 77,18±14,9), AD(83,94±11,4, 90,27±10,0 e 71,98±19,8) e NB(60,89±10,4, 80,58±11,95 e 59,69±12,2) para as combinações vitória x empate, vitória x derrota e empate x derrota. Tais valores são apresentados na tabela 6.31. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.21.

Tabela 6.31: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010 com as características reduzidas.

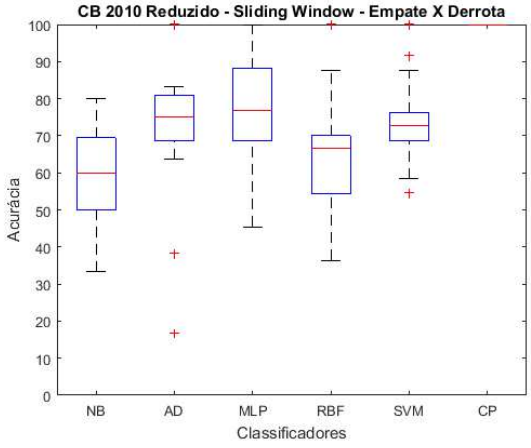
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	60,89±10,4	80,58±11,95	59,69±12,2
AD	83,94±11,4	90,27±10,0	71,98±19,8
MLP	83,45±12,4	92,95±6,6	77,18±14,9
RBF	69,29±10,9	85,85±9,6	64,89±14,7
SVM	77,97±11,3	91,39±5,6	73,48±11,5
CP	98,11±4,8	99,15±2,3	98,46±3,4



(a)



(b)



(c)

Figura 6.21: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2010 com as características reduzidas.

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.31, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(98,11±4,8), AD(83,94±11,4), MLP(83,45±12,4), SVM(77,97±11,3), RBF(69,29±10,9) e NB(60,89±10,4). Através do gráfico 6.21a podemos observar que a ordem de desempenho foi CP, AD, MLP, SVM, RBF e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(99,15±2,3), MLP(92,95±6,6), SVM(91,39±5,6), AD(90,27±10,0), RBF(85,85±9,6) e NB(80,58±11,95). Através do gráfico 6.21b podemos observar que a ordem de desempenho foi CP, MLP, SVM, AD, RBF e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores, nota se ainda uma melhora no desempenho em relação a combinação vitória x empate. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(98,46±3,4), MLP(77,18±14,9), SVM(73,48±11,5), AD(71,98±19,8),RBF(64,89±14,7) e NB(59,69±12,2). Da mesma maneira através do gráfico 6.21c podemos observar que a ordem de desempenho foi CP, AD, MLP, SVM, RBF e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores.

Em relação a comparação entre o conjunto original de 54 características para a redução de 18 a análise será realizada de três maneiras distintas: uma tabela 6.29 contendo a média da acurácia das três combinações (vitória x empate, vitória x derrota, empate x derrota), uma tabela contendo a varição da acurácia 6.30, diferença entre o valor base e o valor obtido por cada um dos três seletores de características(CP, PCA e Relief), e uma figura **??** contendo um gráfico de barras com variação da acurácia distribuída pelos classificadores.

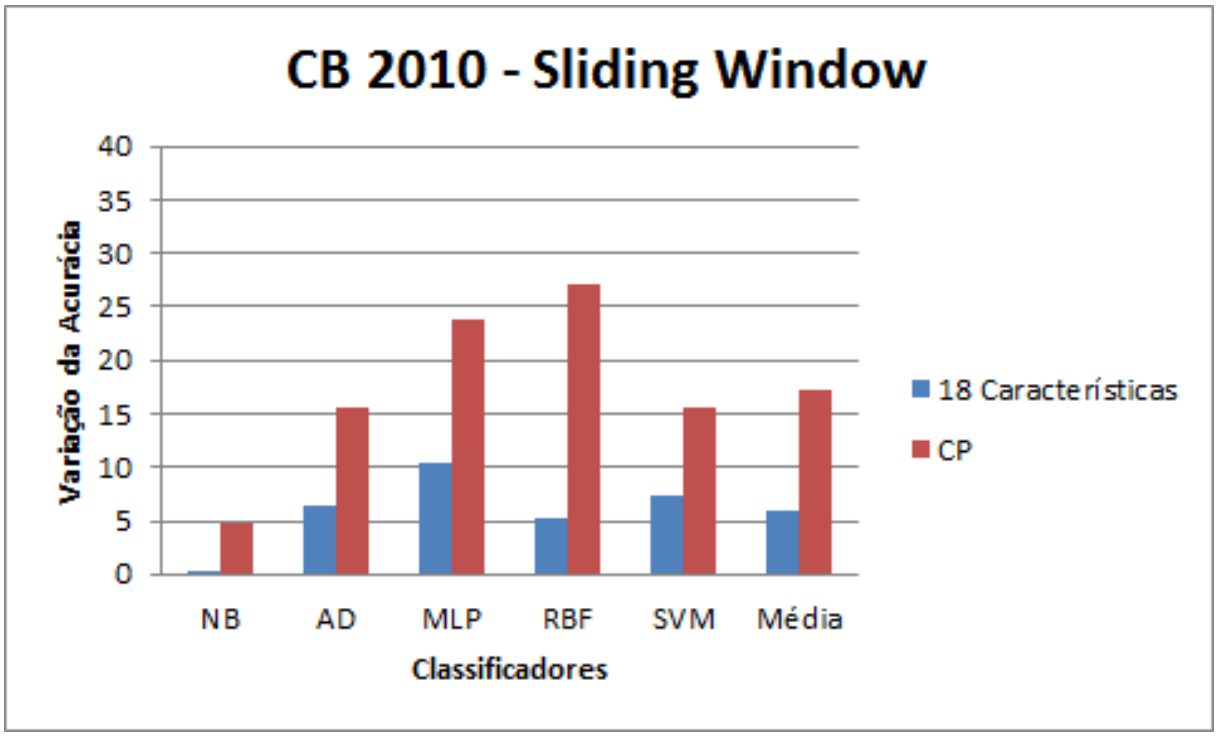


Figura 6.22: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica sliding window para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

Tabela 6.32: Resultados obtidos pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica cross validation.

	NB	AD	MLP	RBF	SVM	Média
54 Características	66,5	76,6	71,8	61,1	71,9	70,1
18 Características	70,93	82,96	81,82	74,78	79,63	78,03
CP	72,1	92,7	97,2	95,2	88,3	89,1

Tabela 6.33: Variação das acurácias obtidas pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica cross validation.

	NB	AD	MLP	RBF	SVM	Média
18 Características	4,43	6,36	10,02	13,68	7,73	7,93
CP	5,6	16,1	25,4	34,1	16,4	19

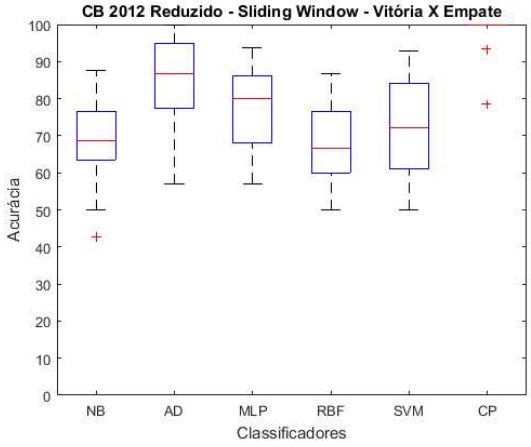
No campeonato CB 2010 com a técnica *slinding window* foram obtidos os seguintes resultados: no classificador NB a acurácia obtida foi de 66,5 com as 54 características originais, 70,93 para as 18 características, aumentando em 4,43 e foi de 72,1 com o CP, aumentando em 5,6. Com a AD a acurácia obtida foi de 76,6 com as 54 características originais, 82,96 para as 18 características, aumentando em 6,36 e foi de 92,7 com o CP, aumentando em 16,1. Para o MLP a acurácia obtida foi de 71,8 com as 54 características originais, 82,96 para as 18 características, aumentando em 10.02 e foi de 97,2 com o CP, aumentando em 25,4. Com o RBF a acurácia obtida foi de 61,1 com as 54 características originais, 74,78 para as 18 características, aumentando em 13,68 e foi de 95,2 com o CP, aumentando em 34,1. Para o SVM a acurácia obtida foi de 70,1 com as 54 características originais, 78,03 para as 18 características, aumentando em 7,73 e foi de 89,1 com o CP, aumentando em 19. O ganho na acurácia média para as 18 características foi de 7,93 e de 19 para o CP. Os valores da acurácia média estão apresentados na tabela 6.32 , a variância da acurácia está apresentada na tabela 6.33 e representados na figura 6.22. Analisando os resultados obtidos pode se afirmar que para essa base de dados na técnica *cross validation* a redução para 18 características melhorou o desempenho para todos os caso mas não superou o desempenho do CP.

A base do campeonato brasileiro 2012 com as características reduzidas apresentou as seguintes acurácias: CP(98,34±5,3, 98,3±4,3 e 98,5±4,2) , SVM(72,19±13,0, 91,74±8,5 e 74,94±14,5), RBF(68,01±10,7, 89,60±7,8 e 66,70±14,1), MLP(78,50±11,8, 91,66±8,6 e 75,28±13,8), AD(84,60±13,6, 92,35±11,9 e 71,93±17,8) e NB(68,20±11,0, 89,02±7,5 e 55,56±12,4) para as combinações vitória x empate, vitória x derrota e empate x derrota.

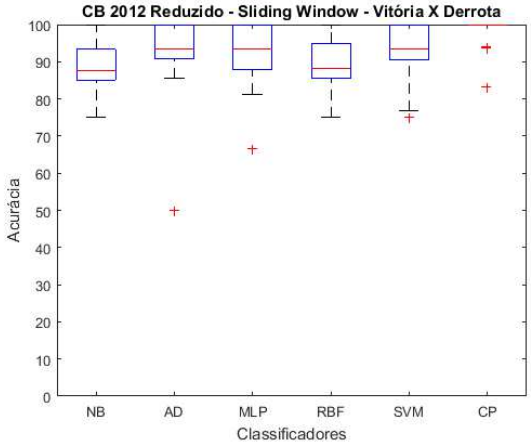
Tais valores são apresentados na tabela 6.34. Os mesmos resultados deram origem aos gráficos *boxplot* apresentados na figura 6.23.

Tabela 6.34: Resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

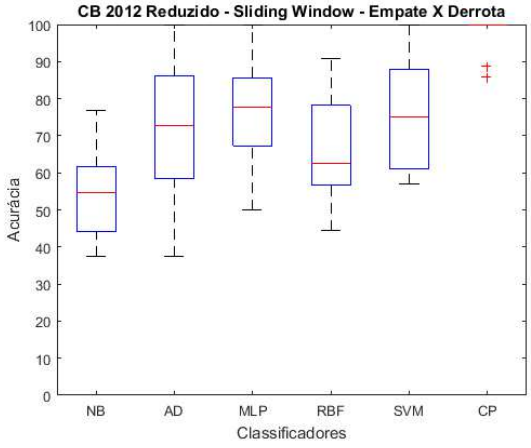
	Vitória X Empate	Vitória x Derrota	Empate x Derrota
NB	68,20±11,0	89,02±7,5	55,56±12,4
AD	84,60±13,6	92,35±11,9	71,93±17,8
MLP	78,50±11,8	91,66±8,6	75,28±13,8
RBF	68,01±10,7	89,60±7,8	66,70±14,1
SVM	72,19±13,0	91,74±8,5	74,94±14,5
CP	98,34±5,3	98,3±4,3	98,5±4,2



(a)



(b)



(c)

Figura 6.23: Gráfico *boxplot* contendo os resultados obtidos pelos classificadores utilizando a técnica Slinding Window para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas..

Analisando os resultados obtidos é possível afirmar que em relação as médias das acurácias, tabela 6.34, obtidas para a combinação vitória x empate a ordem de desempenho foi: CP(98,34±5,3), AD(84,60±13,6), MLP(78,50±11,8), SVM(72,19±13,0), NB(68,20±11,0) e RBF(68,01±10,7). Através do gráfico 6.23a podemos observar que a ordem de desempenho foi: CP, AD, MLP, SVM, NB e RBF com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores. Para a combinação vitória x derrota em ordem das acurácias temos a seguinte ordem: CP(98,3±4,3), AD(92,35±11,9), SVM(91,74±8,5), MLP(91,66±8,6), RBF(89,60±7,8) e NB(89,02±7,5). Através do gráfico 6.23b podemos observar que a ordem de desempenho foi: CP, AD, SVM, MLP, NB e RBF com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores. No caso da combinação empate x derrota as melhores acurácias obtidas foram: CP(98,5±4,2), MLP(75,28±13,8), SVM(74,94±14,5), AD(71,93±17,8), RBF(66,70±14,1) e NB(55,56±12,4). Através do gráfico 6.23b podemos observar que a ordem de desempenho foi: CP, MLP, SVM, AD, RBF e NB com performances bem distintas em relação a mediana, quartis, valores discrepantes e limites superiores e inferiores.

No campeonato CB 2012 com a técnica *sliding window* foram obtidos os seguintes resultados: no classificador NB a acurácia obtida foi de 67 com as 54 características originais, 67,06 para as 18 características, aumentando em 0,06 e foi de 71,7 com o CP, aumentando em 4,7. Com a AD a acurácia obtida foi de 75,7 com as 54 características originais, 82,07 para as 18 características, aumentando em 6,37 e foi de 91,2 com o CP, aumentando em 15,5. Para o MLP a acurácia obtida foi de 74 com as 54 características originais, 84,53 para as 18 características, aumentando em 10,53 e foi de 97,8 com o CP, aumentando em 23,8. Com o RBF a acurácia obtida foi de 68,2 com as 54 características originais, 73,35 para as 18 características, aumentando em 5,15 e foi de 95,4 com o CP, aumentando em 27,2. Para o SVM a acurácia obtida foi de 73,6 com as 54 características originais, 80,95 para as 18 características, aumentando em 7,35 e foi de 89,2 com o CP, aumentando em 15,6. O ganho na acurácia média para as 18 características foi de 5,89 e de 17,2 para o CP. Os valores da acurácia média estão apresentados na tabela 6.35 , a variância da acurácia está apresentada na tabela 6.36 e representados na figura 6.24. Analisando os resultados obtidos

pode se afirmar que para essa base de dados na técnica *sliding window* a redução para 18 características melhorou o desempenho para todos os caso mas não superou o desempenho do CP.

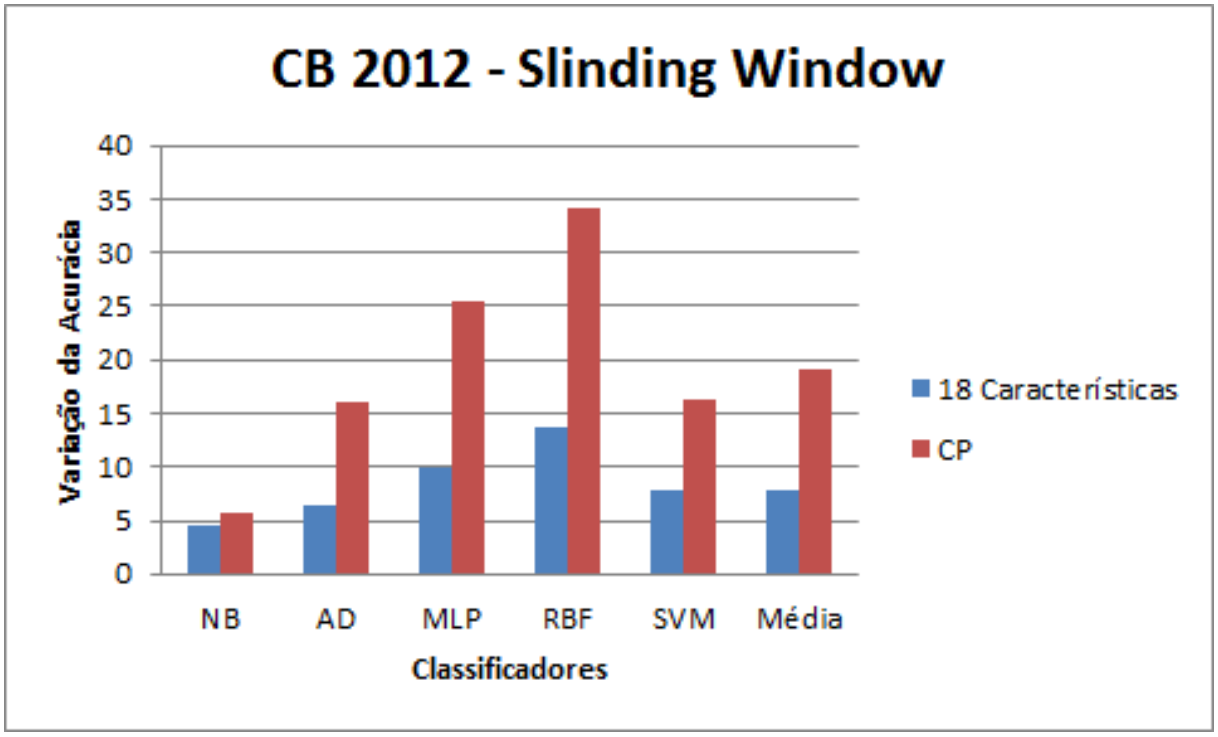


Figura 6.24: Gráfico contendo a variação das acurácias obtidas pelos classificadores utilizando a técnica Cross Validation para os dados referentes ao Campeonato Brasileiro de 2012 com as características reduzidas.

Tabela 6.35: Resultados obtidos pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica slinding window.

	NB	AD	MLP	RBF	SVM	Média
54 Características	67	75,7	74	68,2	73,6	71,7
18 Características	67,06	82,07	84,53	73,35	80,95	77,59
CP	71,7	91,2	97,8	95,4	89,2	88,9

Tabela 6.36: Variação das acurácias obtidas pelos classificadores utilizando a redução para 18 características para o CB 2012 com a técnica slinding window.

	NB	AD	MLP	RBF	SVM	Média
18 Características	0,06	6,37	10,53	5,15	7,35	5,89
CP	4,7	15,5	23,8	27,2	15,6	17,2

Analisando os resultados obtidos pode se afirmar que para essa base de dados na técnica *cross validation* a redução para 18 características melhorou o desempenho para todos os caso mas não superou o desempenho do CP.

6.5 Testes Estatísticos

6.5.1 Teste T-Student

O teste t de Student foi usado para cada característica das partidas de futebol e os valores de p obtidos de cada uma podem ser usados como uma medida do quanto são efetivas na separação dos grupos. Os resultados obtidos estão apresentados nas figuras 6.25, 6.26, 6.27 e 6.28 . As figuras apresentam a função de distribuição acumulada empírica (DCE) dos valores de p para os conjuntos de dados investigados CI 2014/15, CE 2014/15, CB2010 e CB2012.

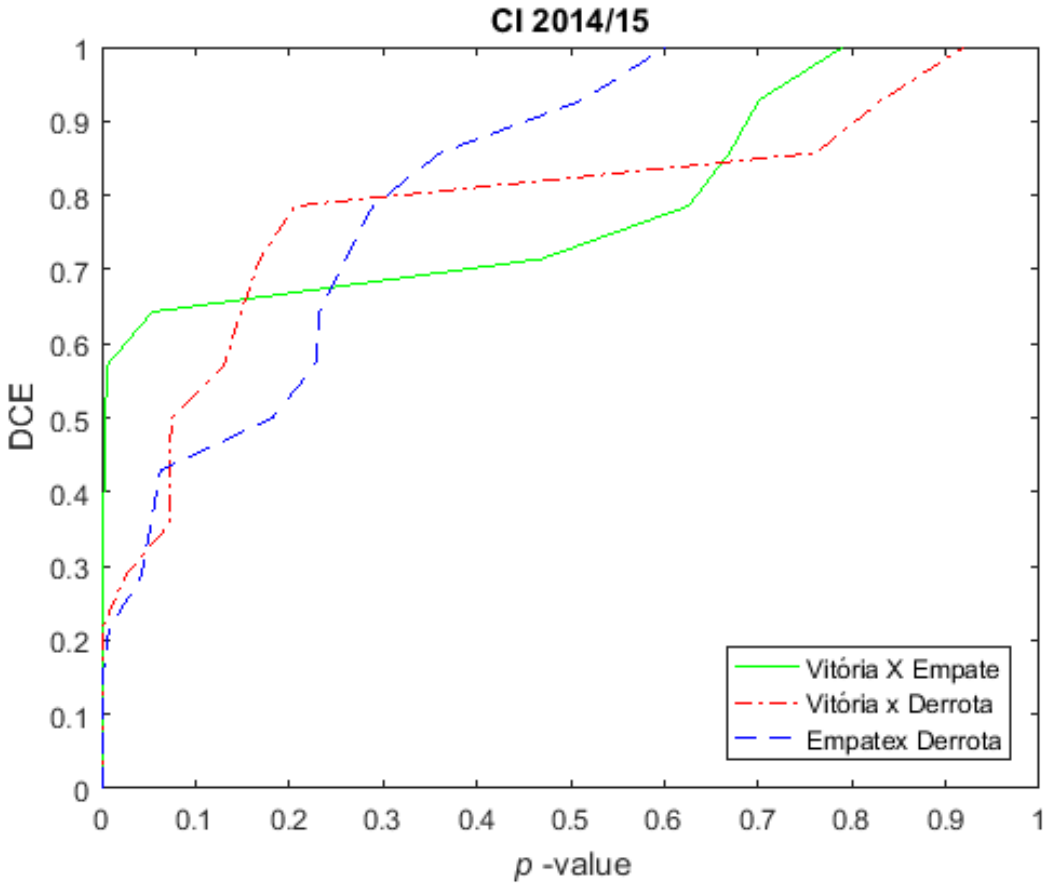


Figura 6.25: Distribuição Acumulada Empírica em função dos p-valores da base de dados CI 2014/15.

Analisando a figura 6.25, que leva em consideração a base de dados com CI 2014/15. É possível observar que para a combinação vitória x empate mais de 50% das características têm p-valores menores que 0,05. Para a combinação vitória x derrota mais de 25% das características têm p-valores menores que 0,05. Enquanto que no caso da combinação empate x derrota mais de 20% das características têm p-valores menores que 0,05. Sendo assim segundo o critério estabelecido por Kunwar [2] pode se afirmar que as características possuem relevância estatística.

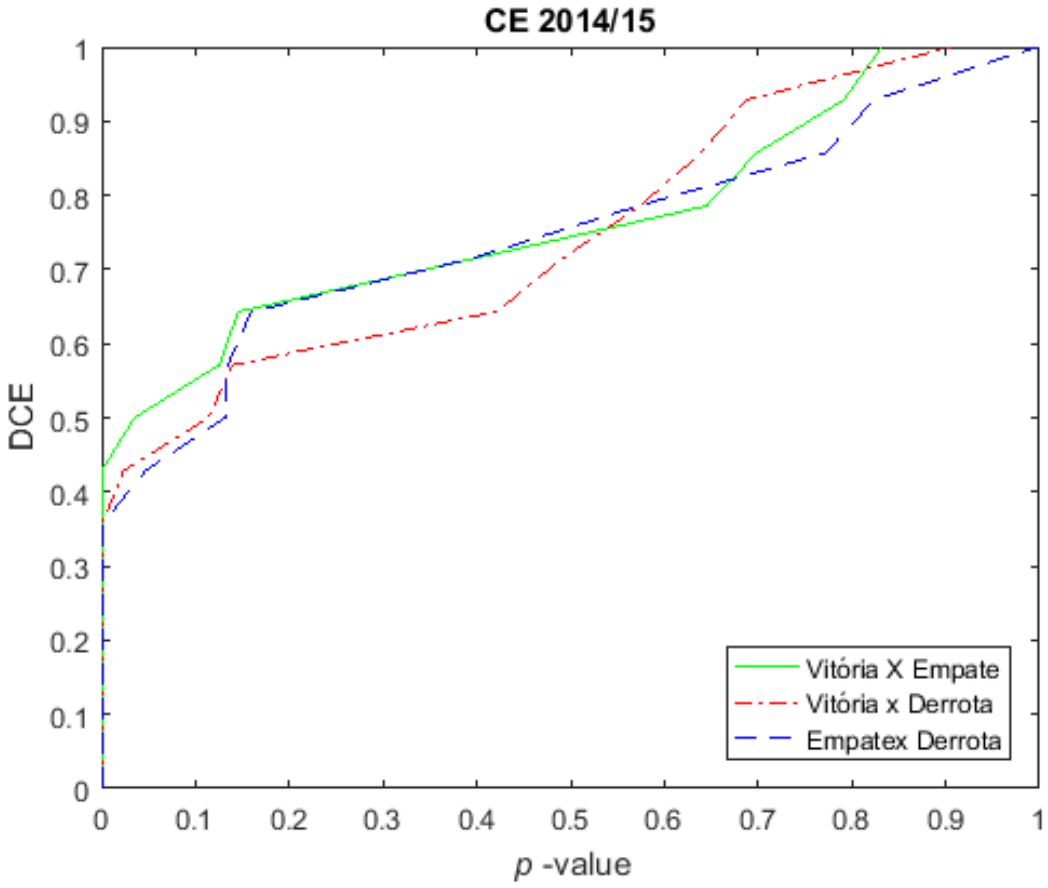


Figura 6.26: Distribuição Acumulada Empírica em função dos p-valores da base de dados CE 2014/15.

Analisando a figura 6.26, que leva em consideração a base de dados com CE 2014/15. É possível observar que para a combinação vitória x empate mais de 43% das características têm p-valores menores que 0,05. Para a combinação vitória x derrota mais de 40% das características têm p-valores menores que 0,05. Enquanto que no caso da combinação empate x derrota mais de 35% das características têm p-valores menores que 0,05. Sendo assim segundo o critério estabelecido por Kunwar [2] pode se afirmar que as características possuem relevância estatística.

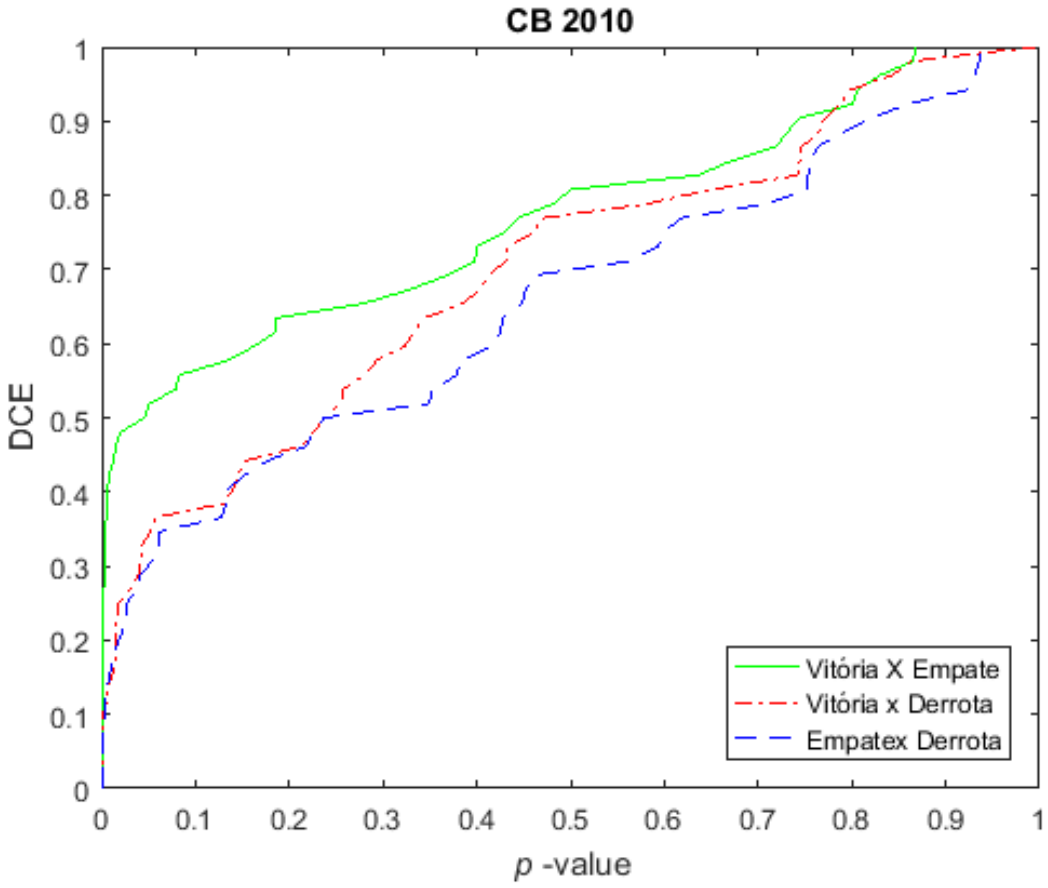


Figura 6.27: Distribuição Acumulada Empírica em função dos p-valores da base de dados CB 2010.

Analisando a figura 6.28, que leva em consideração a base de dados com CB 2010. É possível observar que para a combinação vitória x empate mais de 45% das características têm p-valores menores que 0,05. Para a combinação vitória x derrota mais de 25% das características têm p-valores menores que 0,05. Enquanto que no caso da combinação empate x derrota mais de 27% das características têm p-valores menores que 0,05. Sendo assim segundo o critério estabelecido por Kunwar [2] pode se afirmar que as características possuem relevância estatística.

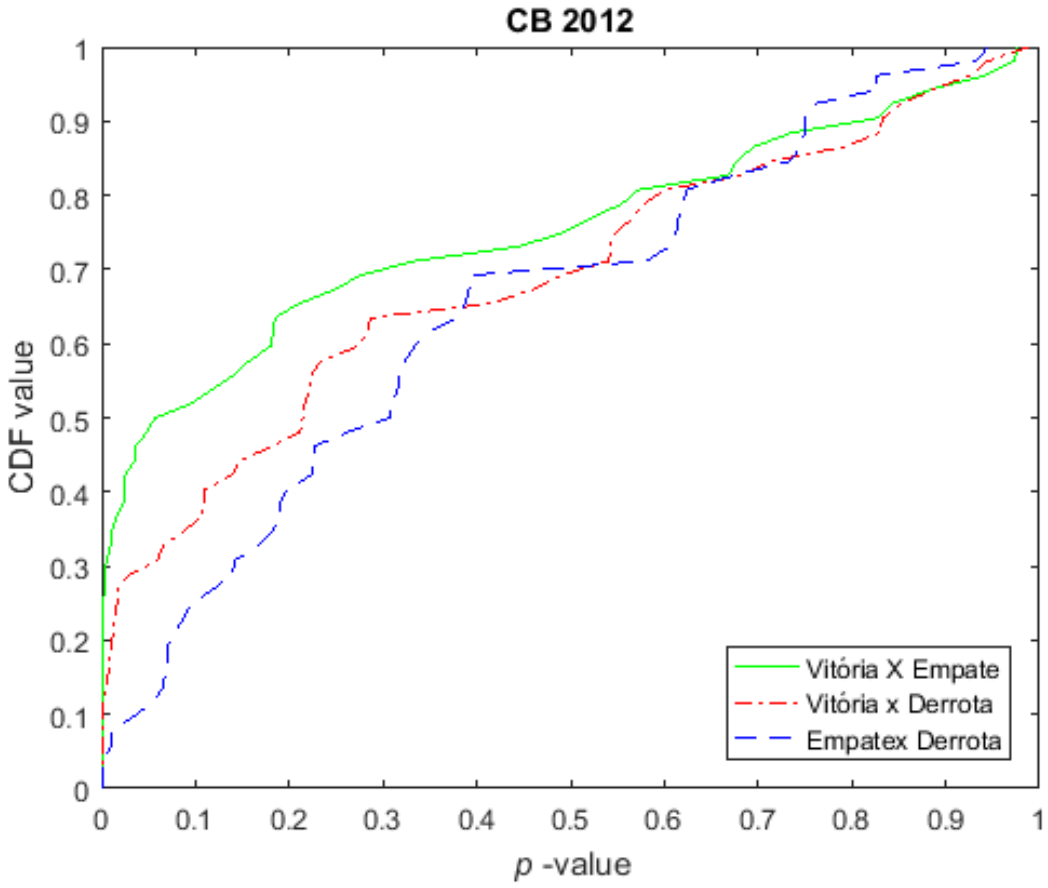


Figura 6.28: Distribuição Acumulada Empírica em função dos p-valores da base de dados CB 2012.

Analisando a figura 6.27, que leva em consideração a base de dados com CB 2012. É possível observar que para a combinação vitória x empate mais de 40% das características têm p-valores menores que 0,05. Para a combinação vitória x derrota mais de 30% das características têm p-valores menores que 0,05. Apenas no caso da combinação empate x derrota temos em torno de 10% das características com p-valores menores que 0,05. Sendo assim segundo o critério estabelecido por Kunwar [2] pode se afirmar que as características possuem relevância estatística.

6.5.2 Rank de Friedman

O Rank de Friedman foi aplicado para os resultados obtidos pelos seis classificadores (CP, NB, AD, MLP, RBF e SVM) nas quatro base de dados testadas (CI 2014/15, CE 2014/15, CB 2010 e CB 2012) para as duas técnicas de validação do método (*cross validation* e *slinding window*). Os resultados foram divididos em quatro tabelas. Duas tabelas contendo os valores da acurácia média calculados sobre os dados (vitória x empate, vitória x derrota e empate x derrota) para cada um dos conjuntos. Para o caso do *cross validation* os resultados se encontram na tabela 6.37, enquanto que para a técnica *sliding window* os valores são mostrados na tabela 6.38. Outras duas tabelas contendo os valores não ajustados e ajustados de p para os procedimentos de Nemenyi, Holm, Shaffer e Bergmann-Hommel para as 15 combinações envolvendo a combinação dos seis classificadores, comparados dois a dois, para o caso da técnica *cross validation* esses valores estão disponíveis na tabela 6.39 e no caso da técnica *sliding window* na tabela 6.40.

Analisando os resultados apresentados na tabela 6.37, que foram obtidos através do procedimento de teste de significância de [25]. Podemos afirmar que o ranking dos classificadores, do melhor para o pior, é CP, MLP, AD, SVM, RBF e NB, considerando os quatro campeonatos e a técnica de validação *cross validation*. O valor de p encontrado para esse caso foi 0,0031. O que pode indicar que existem diferenças estatísticas significativas entre os métodos, de acordo com [15], que define valores abaixo de 0,05 para rejeitar a hipótese nula.

.

Tabela 6.37: Acurácia Média obtida pelos classificadores com a técnica do cross validation para as 4 base de dados

	Classificadores					
Base de Dados	CP	NB	AD	MLP	RBF	SVM
CI 2014/15	99,7	80,7	99,4	100,0	97,4	98,6
CE 2014/15	99,5	83,1	99,4	100,0	97,3	93,4
CB 2010	99,6	79,0	97,1	94,9	82,7	97,6
CB 2012	98,9	82,0	97,0	95,2	84,3	97,2
Rank de Friedman	1,5	6,0	3,0	2,5	4,75	3,25
p-value	0,0103					

Analisando os resultados apresentados na tabela 6.37, que foram obtidos através do procedimento de teste de significância de [25]. Podemos afirmar que o ranking dos classificadores, do melhor para o pior, é CP, AD, MLP, SVM, RBF e NB, considerando os quatro campeonatos e a técnica de validação *cross validation*. O valor de p encontrado para esse caso foi 0,0103. O que pode indicar que existem diferenças estatísticas significativas entre os métodos, de acordo com [15].

Tabela 6.38: Acurácia Média obtida pelos classificadores com a técnica do sliding window para as 4 base de dados

	Classificadores					
Base de Dados	CP	NB	AD	MLP	RBF	SVM
CI 2014/15	97,7	70,3	85,6	87,8	78,1	79,3
CE 2014/15	97,5	75,1	86,1	84,5	79,3	78,6
CB 2010	99,8	67,0	75,7	74,0	68,2	73,6
CB 2012	99,8	66,5	76,6	71,8	64,1	71,9
Rank de Friedman	1,0	5,75	2,25	3,0	5,0	4,0
p-value	0,0031					

Analisando os resultados apresentados na tabela 6.39 sob a ótica de Trawinski [64], que afirma que se o valor ajustado p-valor para uma hipótese nula individual for menor que 0,05, essa hipótese é rejeitada. Temos então que para a técnica *cross validation* apenas 5 dos 15 pares não possuem diferenças estatísticas significativas: CP *versus* NB, NB *versus* MLP, CP *versus* RBF, NB *versus* AD e NB *Versus* SVM. Com os procedimentos Shaffers e Bergmann-Hommels, apenas o par CP *versus* NB foi rejeitado. Todos esses casos estão destacados na tabela em itálico.

Tabela 6.39: Valores ajustados p por 2 *versus* 2 comparações para as 15 hipóteses para cada classifcador com a técnica *cross validation*.

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	CP vs .NB	*.6972E-4*	*0.01004*	*0.01004*	*0.01004*	*0.01004*
2	NB vs .MLP	*0.00815*	0.12226	0.11411	0.08150	0.08150
3	CP vs .RBF	*0.01401*	0.21028	0.18225	0.14019	0.14019
4	NB vs .AD	*0.02334*	0.35013	0.28010	0.23342	0.16339
5	NB vs .SVM	*0.037635*	0.56452	0.41398	0.37635	0.225811
6	MLP vs .RBF	0.08897	1.33459	0.88973	0.889730	0.53383
7	CP vs .SVM	0.18587	2.78815	1.67289	1.30113	1.30113
8	AD vs .RBF	0.18587	2.78815	1.67289	1.30113	1.30113
9	CP vs .AD	0.25683	3.85258	1.79787	1.79787	1.30113
10	RBF vs .SVM	0.25683	3.85258	1.79787	1.79787	1.30113
11	NB vs .RBF	0.34470	5.17056	1.79787	1.79787	1.37881
12	CP vs .MLP	0.44969	6.74537	1.79876	1.79876	1.37881
13	MLP vs .SVM	0.57075	8.56125	1.79876	1.79876	1.71225
14	AD vs .MLP	0.70545	10.58185	1.79876	1.79876	1.71225
15	AD vs .SVM	0.85010	12.75160	1.79876	1.79876	1.71225

Analisando os resultados apresentados na tabela 6.40 ainda sob a ótica de Trawinski [64]. Temos então que para a técnica *sliding window* apenas 6 dos 15 pares não possuem

diferenças estatísticas significativas: CP *versus* NB, NB *versus* MLP, CP *versus* RBF, NB *versus* AD, NB *Versus* SVM e CP *versus* SVM. Com os procedimentos Shaffers e Bergmann-Hommels, apenas dois pares CP *versus* NB e CP *versus* RBF foi rejeitado. Todos esses casos estão destacados na tabela em itálico.

Tabela 6.40: Valores ajustados p por 2 *versus* 2 comparações para as 15 hipóteses para cada classifcador com a técnica *sliding window*.

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	CP vs .NB	*3.2983E-4*	*0.00494*	*0.00494*	*0.00494*	*0.00494*
2	CP vs .RBF	*0.00249*	*0.03745*	*0.03495*	*0.02496*	*0.02496*
3	NB vs .AD	*0.00815*	0.12226	0.10596	0.081509	0.08150
4	CP vs .SVM	*0.02334*	0.35013	0.28010	0.23342	0.16339
5	NB vs .MLP	*0.03763*	0.56452	0.41398	0.37635	0.26344
6	AD vs .RBF	*0.03763*	0.56452	0.41398	0.37635	0.26344
7	CP vs .MLP	0.13057	1.95855	1.17513	0.91399	0.78342
8	MLP vs .RBF	0.13057	1.95855	1.17513	0.91399	0.78342
9	NB vs .SVM	0.18587	2.78815	1.30113	1.30113	0.78342
10	AD vs .SVM	0.18587	2.78815	1.30113	1.30113	0.78342
11	CP vs .AD	0.34470	5.1705633	1.72352	1.37881	1.03411
12	MLP vs .SVM	0.44969	6.74537	1.79876	1.79876	1.03411
13	RBF vs .SVM	0.44969	6.74537	1.79876	1.79876	1.03411
14	NB vs .RBF	0.57075	8.56125	1.79876	1.79876	1.14150
15	AD vs .MLP	0.57075	8.56125	1.79876	1.79876	1.14150

6.6 Comparação com o Estado da Arte

Na literatura existem uma série de estudos com o objetivo de investigar o desempenho das equipes, bem como prever os resultados de partidas de futebol. Todos esses autores apresentam uma metodologia adequada, embora sua consistência só seja evidenciada pelos

resultados finais.

A tabela 6.41 apresenta vários métodos para a predição de partidas de futebol e seus respectivos valores de acurácia, incluindo os obtidos nesse presente estudo. Em geral, o algoritmo CP obteve resultados relevantes na previsão de partidas de futebol. Nestes testes comparativos, cada autor escolheu uma metodologia específica e promissora, no entanto, não podemos compará-los para definir o melhor, pois esta seria uma tarefa difícil, para não dizer sem sentido. Na verdade, as diferentes metodologias são muito mais complementares do que classificáveis. Portanto, é possível afirmar que o método aqui proposto forneceu a robustez desejada para aplicações semelhantes.

6.7 Considerações Finais Deste Capítulo

Este capítulo apresentou os resultados obtidos de acordo com a metodologia proposta. É possível afirmar que nos resultados apresentados na Seção 6.2 o CP obteve as melhores acurácias quando comparados com os outros cinco classificadores utilizados: RB, AD, MLP, RBF e SVM. Também é possível afirmar que pelos resultados apresentados na Seção 6.3 que o CP conseguiu melhorar a acurácia dos cinco classificadores com índices superiores ao Relief e ao PCA. Com a seção 6.5 é possível garantir a relevância estáticas das características e dos classificadores. Já a seção 6.4 demonstra que reduzir as características coletadas pelo método *scout* também pode melhorar as acurácias dos classificadores. De acordo com os resultados apresentados na seção 6.6 as acurácias obtidas com essa metodologia são tão boas ou superiores aos resultados encontrados no estado da arte, variando de 0,96 a 0,99.

Tabela 6.41: Comparação com o estado da arte

Estudo	Base de Dados	Características	Métodos	Acurácia
[67]	Campeonato Inglês	Fundamentos	Baseline, NaiveBayes, HMM, MNB, RBF, SVM, RF, Linear SVM, One vs ALL SGD	0,52
[11]	Liga dos Campeões da UEFA	Fundamentos	NB, K-NN, ANN, BayssianNet, Logoost, RF, ANN	0,68
[49]	Campeonato Espanhol	Fundamentos Dados fisiológicos	NETICA	0,92
[34]	Campeonato Inglês	Fundamentos	SVM	0,53
[62]	Campeonato Holândes	Fundamentos Histórico	HIRP,DTN, LogitBooST, FURIA, J48, HyperPipes, MP NaiveBayes, RF	0,56
[20]	Campeonato Português	Histórico Fundamentos Dados fisiológicos	C5.0, JRip, RF KNN, SVM, NB	0,58
Esta Tese	CI - 2014/15 CE - 2014/15 CB - 2010 CB - 2012	Fundamentos	CP	0,99 0,99 0,99 0,98

Capítulo 7

Conclusão

7.1 Introdução

Este capítulo apresenta as conclusões e as contribuições deste trabalho, a publicação e os trabalhos futuros que poderão ser desenvolvidos a partir desta tese e sua publicação.

7.2 Conclusões

O objetivo deste trabalho é explorar o método Classificador Polinomial como ferramenta de predição de resultados de partidas de futebol. Esse classificador mapeia um conjunto de amostras de treinamento e classifica as amostras previamente identificadas.

Os seguintes objetivos específicos deste trabalho foram alcançados:

- mostrar como o classificador polinomial se comporta na classificação de padrões de quatro bases de dados distintas envolvendo partidas de futebol a metodologia apresentada no capítulo 5 tem como um dos objetivos demonstrar esse comportamento;
- comparar e demostrar a maior eficácia do classificador polinomial com a de outros classificadores utilizados no estado da arte de predição de resultados de partidas de futebol conforme podemos observar nos resultados obtidos na seção 6.2;

- analisar o classificador polinomial como um método de otimização utilizado na seleção das melhores características para predição de resultados de partidas de futebol a metodologia apresentada no capítulo 5 tem como um dos objetivos demonstrar esse comportamento; e

- comparar e demonstrar a maior eficácia do classificador polinomial com a de outros seletores de características utilizados no estado da arte de predição de resultados de partidas de futebol conforme podemos observar nos resultados obtidos na seção 6.3.

7.3 Contribuições Deste Trabalho

Este trabalho apresenta o classificador polinomial como ferramenta de predição de resultados das partidas de futebol. Foram realizados testes considerando o Classificador Polinomial como um algoritmo de classificação e também como de seleção de características.

O classificador polinomial enquanto algoritmo de classificação conseguiu resultados superiores a outros cinco classificadores: Naive Bayes, Árvore de Decisão, MLP, RBF e SVM. Os resultados foram superiores também a de outros trabalhos publicados na área mas que levam em conta distintas metodologias e tipos de dados.

O Classificador Polinomial enquanto algoritmo de seleção de características foi comparado com outras duas técnicas de seleção de características: Análise de Componentes Principais e Relif. Mostrando que esse primeiro classificador apresenta melhores resultados.

7.4 Publicação Deste Trabalho

- R. G. Martins, A.S. Martins, L. A. Neves, L.V. Lima, E. L. Flôres and M. Z. Nascimento, Exploring polynomial classifier to predict match results in football championships, Experts Systems With Applications, 83, 79-93, 2017.

7.5 Trabalhos Futuros

Este trabalho possibilita pesquisas em classificadores polinomiais como algoritmo de seleção de características aplicadas a novas áreas de estudos como por exemplo classificação de imagens médicas.

Do ponto de vista de partidas de futebol muitas novas possibilidades se abrem:

- Exploração do uso do classificador polinomial na previsão de resultados de partidas de futebol;
- Utilização de dados espaço temporal na predição de resultados de partidas de futebol;
- Exploração do uso do classificador polinomial na previsão de resultados de partidas de outras modalidades esportivas tais como: basquete, vôlei, hadebol, etc;

7.6 Considerações Finais Deste Capítulo

Outros estudos foram desenvolvidos para prever o resultado dos resultados de partidas de futebol utilizando algoritmos de aprendizagem de máquinas. Esses algoritmos são ferramentas que recebem como entrada um conjunto de características e fornece como saída a previsão do resultados (vitória, empate e derrota). Existem algoritmos que podem fornecer uma resposta mais adequada ao problema [51, 55]. A decisão da utilização de aplicar o classificador polinomial para fazer a predição de resultados de jogos de futebol é baseada na sua capacidade de aprendizagem complexa. Pois ele é capaz de atuar em padrões que podem ser linearmente inseparáveis e o sucesso obtido em outras aplicações [50]. O classificador polinomial utiliza parametrização não linear que expande de maneira não linear uma sequência de vetores de entrada para uma dimensão superior e mapeia-os para uma sequência de saída desejada. Essa expansão pode melhorar a separação das diferentes classes em um espaço vetorial. Além disso, essa estratégia apresenta as vantagens de fornecer apenas um modelo para separação ótima das classes e dessa maneira pode solucionar o problema o que não ocorre com os modelos apresentados em outros trabalhos [10, 53].

Este capítulo apresentou as conclusões, contribuições e a publicação deste trabalho, os trabalhos futuros que poderão ser desenvolvidos a partir desta tese.

Referências Bibliográficas

[1] Balduck A. L.; Prinzie A.; Buelens M. , The effectiveness of coach turnover and the effect on home team advantage, team quality and team ranking, Journal of Applied Statistics, vol. 37, no. 4, pp. 679?689, 2010.

[2] Bhatia, K. S. S.; ; Lam, A. C. L.; Pang, S. W. A.; Wang, D.; Ahuja, A. T., Feasibility Study of Texture Analysis Using Ultrasound Shear Wave Elastography to Predict Malignancy in Thyroid Nodules, Ultrasound in Medicine Biology, 42, 7, 1671 - 1680, 2016.

[3] Beck, N., Meyer, M., Modeling team performance, Empirical Economics, 2012, 43, 1, 335–356, 1435-8921, 10.1007/s00181-011-0463-2, http://dx.doi.org/10.1007/s00181-011-0463-2.

[4] Bhandari, I. and Colet, E., Parker, J., Pines, Z., Pratap, R.,Ramanujam, K.,Advanced Scout: Data Mining and Knowledge Discovery in NBA Data, Data Mining and Knowledge Discovery, 1997, 1, 1, 121–125, 1573-756X, 10.1023/A:1009782106822, http://dx.doi.org/10.1023/A:1009782106822.

[5] Bittner E. ; Nußbaumer A.; Janke W. ; Weigel M., Self-affirmation model for football goal distributions, EPL (Europhysics Letters),78,5,58002, http://stacks.iop.org/0295-5075/78/i=5/a=58002,2007.

[6] Brooks, J., Kerr, M., Guttag, J., 2016. Using machine learning to draw inferences from pass location data in soccer. Statistical Analysis and Data Mining: The ASA Data Science Journal 9 (5), 338?349.

[7] D. S. Broomhead, D. L., 1988. Radial basis functions, multi-variable func- tional interpolation and adaptive networks 2 (3), 321-355.

[8] Bruno, D. O. T., do Nascimento, M. Z., Ramos, R. P., Batista, V. R., Neves, L. A., Martins, A. S., 2016. Lbp operators on curvelet coe?cients as an algorithm to describe texture in breast cancer tissues. Expert Systems with Applications 55, 329-340.

[9] Caliman, G. B.; Ferreira, R. B. 2006. Uma proposta de ?scout? tático para o futebol. 2006. Monografia. Faculdade Salesiana de Vitória, Vitória.

[10] Campbell, W. M., Assaleh, K. T., Broun, C. C., 2002. Speaker recognition with polynomial classifiers. Speech and Audio Processing, IEEE Transac- tions on 10 (4), 205-212.

[11] Constantinou, A. C., Fenton, N. E., Neil, M., 2013. Pro?ting from an inef- ?cient association football gambling market: Prediction, risk and uncer- tainty using bayesian networks. Knowledge-Based Systems 50, 60-86.

[12] Cortes, C., Vapnik, V., Sep. 1995. Support-vector networks. Mach. Learn. 20 (3), 273-297.

[13] Cunha S. A.; Binotto M. R.; Barros L. R. M.,Análise da variabilidade na medição de posicionamento tático no futebol., Revista paulista de educação física,2001, 15, 2, 111-116.

[14] Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2 (4), 303-314.

[15] Demsar J. , Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res., vol. 7, pp. 1?30, Dec. 2006.

[16] De Paola M.; Scoppa V., The effects of managerial turnover evidence from coach dismissals in italian soccer teams, Journal of Sports Economics, vol. 13, no. 2, pp. 152-168, 2012.

[17] Do Nascimento, M. Z., Martins, A. S., Neves, L. A., Ramos, R. P., Flo- res, E. L., Carrijo, G. A., 2013. Classi?cation of masses in mammographic image using wavelet domain features and polynomial classi?er. Expert Systems with Applications 40 (15), 6213-6221.

[18] Drubscky, R., O universo tático do futebol: escola brasileira., Health, 2003, Segunda Edição, ISBN 0000575763.

[19] Dua, S., Singh, H., Thompson, H., 2009. Associative classi?cation of mam- mograms using weighted rules. Expert Systems with Applications 36 (5), 9250-9259.

[20] Duarte, L., Soares, C., Teixeira, J., 2015. Previs?ao de resultados de jogos de futebol. Master´s thesis, Faculdade da Engenharia da Universidade do Porto.

[21] Duda, R. O.; HART, P. E.; STORK, D. G. Pattern Classification. 2. ed. New York: Wiley-Interscience; 2000.

[22] Fawcett, T., 2006. An introduction to roc analysis. Pattern Recognition Let- ters 27 (8), 861-874, rOC Analysis in Pattern Recognition.

[23] Federation Internationale de Football Association, "Official Documents ? FIFA World Cup", http://www.fifa.com/about-fifa/official-documents/index.html, acessado em 18 de maio de 2016.

[24] Forrest D. ; Goddard J. ; Simmons R. , ?Odds-setters as forecasters: The case of english football, International Journal of Forecasting, vol. 21, no. 3, pp. 551 - 564, 2005.

[25] Garcia S. , Herrera F., An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons, Journal of Machine Learning Research,9,Dec, 2677–2694, 2008.

[26] Gardner, M.J., Altman, D.G., Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal (Clin Res Ed) 1986;292(6522):746-50.

[27] Gutierrez O, Ruiz J.L., Game Performance Versus Competitive Performance in the World Championship of Handball 2011., Journal of Human Kinetics., 2013, 36, 137–147, 10.2478/hukin-2013-0014, http://doi.org/10.2478/hukin-2013-0014.

[28] Hall M. , Frank E., Holmes G., Pfahringer, B., Reutemann, P., Witten, I. H., Nov. 2009. The weka data mining software: An update. SIGKDD Explor. Newsl. 11 (1), 10?18.

[29] Hall, M. A., 2000. Correlation-based feature selection for discrete and nu- meric class machine learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. ICML ?00. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 359-366.

[30] Heuer A.; Muller C.; Rubner O.; Hagemann N.; Strauss B., Usefulness of dismissing and changing the coach in professional soccer, PloS one, vol. 6, no. 3, p. e17664, 2011.

[31] Huan K.; Chang W., A Neural Network Method for Prediction of 2006 World Cup Football Game., 978, 1, 8126, 2010.

[32] Hucaljuk, J., Rakipovi c, A., 2011. Predicting football scores using machine learning techniques. In: MIPRO, 2011 Proceedings of the 34th Interna- tional Convention. IEEE, pp. 1623?1627.

[33] Hunter, D., New Rules and Fancies, New Statesman, 3 Jan. 2011, 57, 2011, 57. Academic OneFile.

[34] Igiri, C. P., 2015. Support vector machinebased prediction system for a football match result. IOSR Journal of Computer Engineering (IOSR-JCE) 17 (3), 21-26.

[35] Janke W.; E. Bittner; A. Nußbaumer; Weigel M., 1607324X, Condensed Matter Physics, 4, 739 - 752, Football fever: self-affirmation model for goal distributions, 12,2009.

[36] John, G. H., Langley, P. Estimating Continous Distributions in Bayesian Classifiers, Proceedings of the eleventh conference on uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, 1995.

[37] Jollie, I., 2014. Principal Component Analysis. John Wiley e Sons, Ltd.Karabatak, M., 2015. A new classifier for breast cancer detection based on naive bayesian. Measurement 72, 32-36.

[38] Karabatak, M., 2015. A new classifier for breast cancer detection based on naive bayesian. Measurement 72, 32-36.

[39] Kira, K., Rendell, L. A., 1992. A practical approach to feature selection. In: Proceedings of the ninth international workshop on Machine learning. pp. 249-256.

[40] Kononenko, I., 1994. Estimating attributes: analysis and extensions of relief. In: European conference on machine learning. Springer, pp. 171-182.

[41] Lamas, L., Barrera, J., Otranto, G., Ugrinowitsch, C., Invasion team sports: strategy and match modeling.,International Journal of Performance Analisys in Sport, 2014, 14, 1, 307–329, 14748185.

[42] Lima, R. A. F. Estratégias de Seleção de Atributos para detecção de anomalias em transações eletrônicas. Dissertação (Mestrado) - Universidade Federal de Minas Gerais, Belo Horizonte, 2016.

[43] Link, D., Ahmann, J., Modern game observation in beach volleyball based on positional data, Sportwissenschaft, 2013, 43, 1, 1–11, 1868-1069, 10.1007/s12662-013-0282-z, http://dx.doi.org/10.1007/s12662-013-0282-z.

[44] Liu, H., Motoda, H. (1998b). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, MA, USA. ISBN 079238198X.

[45] Martins, A. S. Separação em uma ou mais classes utilizando o classificador polinomial. Tese (Doutorado) - Universidade Federal de Uberlândia, Uberlândia, 2013.

[46] Moura F.A., Santana J.E. and Vieira N.A., Santiago P.R.P., Cunha S.A., Analysis of Soccer Players? Positional Variability During the 2012 UEFA European Championship: A Case Study., Journal of Human Kinetics., 47, 1, 225-236, 2015, 10.1515/hukin-2015-0078, http://doi.org/10.1515/hukin-2015-0078.

[47] Oficial Website of the Olympic Movement, "All Facts - London 2012", http://www.olympic.org/london-2012-summer-olympics, acessado em 18 de maio de 2016.

[48] Owramipur F.; Eskandarian P.; Mozneb F. S. , Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team, International Journal of Computer Theory and Engineering, vol. 5, no. 5, pp. 812:815, 2013. [Online]. http://www.ijcte.org/index.php?m=content&c=index&a=show&catid=51&id=925

[49] Parinaz, O. F. E., Sadat, M. F., 2013. Football result prediction with bayesian network in spanish league-barcelona team. International Journal of Computer Theory and Engineering 5 (5), 812-815.

[50] Park, B. J., Oh, S. K., Kim, H. K., 2008. Design of polynomial neural network classi?er for pattern classi?cation with two classe. Journal of Electrical Engineering e Technology 3 (1), 108-114.

[51] Pendharkar, P., Khosrowpour, M., Rodger, J., 2000. Application of bayesian network classi?ers and data envelopment analysis for mining breast cancer patterns. Journal of Computer Information Systems 40 (4), 127-132.

[52] Perin, C., Vuillemot, R., Fekete, J. D., SoccerStories: A Kick-off for Visual Soccer Analysis, IEEE Transactions on Visualization and Computer Graphics, 2013, 19, 12, 2506-2515, 10.1109/TVCG.2013.192, 1077-2626.

[53] P. K. Ajmera, R. S. H., 2010. Speaker recognition using auditory fea- tures polynomial classi?er. International Journal of Computer Applications 1 (14), 86?91.

[54] Quinlan, J. R., 1986. Induction of decision trees. Machine Learning 1 (1), 81?106.

[55] Ramirez-Villegas, J. F., Ramirez-Moreno, D. F., 2012. Wavelet packet energy, tsallis entropy and statistical parameterization for support vector-based and neural-based classi?cation of mammographic regions. Neurocomputing 77 (1), 82 ? 100.

[56] Riedmiller and H. Braun, A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, Proc. ICNN, San Francisco (1993).

[57] Scholz, M. Approaches to analyse and interpret biological profile data,2006.

[58] Sfeir, M. N., Laws of the Game: (adapted from FIFA 2010-11), World Literature Today, 85.3 (May-June 2011), 38–40, 2011, 38. Academic OneFile.

[59] Silva R.; Vainstein M. H.; Lamb L. C; Prado S. D., A simple non-Markovian computational model of the statistics of soccer leagues: Emergence and scaling effects, Computer Physics Communications, 184,3, 661 - 670, 0010-4655,http://dx.doi.org/10.1016/j.cpc.2012.10.030, 2013.

[60] Silva R.; Dahmen S. R., Universality in the distance between two teams in a football tournament, Physica A: Statistical Mechanics and its Applications, 398, 56 - 64, 0378-4371, http://dx.doi.org/10.1016/j.physa.2013.12.008, 2014.

[61] Streib N., Young S. J., Sokol J., A Major League Baseball Team Uses Operations Research to Improve Draft Preparation, Interfaces, 42, 2, 119-130, 2012, 10.1287/inte.1100.0552, http://dx.doi.org/10.1287/inte.1100.0552.

[62] Tax, N., Joustra, Y., 2015. Predicting the dutch football competition using public data: A machine learning approach. Transactions on Knowledge and Data Engineering 10 (10), 1-13.

[63] Timmaraju A. S. ; Palnitkar A. ; V. Khanna, Game ON! Predicting English Premier League Match Outcomes, 2013.

[64] Trawinski B., Smetek M. , Lasota T., G. Trawinski, Evaluation of Fuzzy System Ensemble Approach to Predict from a Data Stream, Intelligent Information and Database Systems: 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014, Proceedings, Part II,2014, Springer International Publishing, Cham, 137–146, 978-3-319-05458-2.

[65] Triola, M. F., Introdução à estatística, Rio de Janeiro: LTC, 2005.

[66] Tufekci, P., 2016. Prediction of Football Match Results in Turkish Super League Games. Springer International Publishing, Cham, pp. 515-526.

[67] Ulmer, B., Fernandez, M., 2013. Predicting soccer match results in the english premier league. Ph.D. thesis, Stanford.

[68] Vapnik, V. N., 1995. The nature of statistical learning theory. Springer-Verlag, New York, Inc., New York, NY, USA.

[69] Vafaeipour, M., Rahbari, O., Rosen, M. A., Fazelpour, F., Ansarirad, P., 2014. Application of sliding window technique for prediction of wind ve- locity time series. International Journal of Energy and Environmental En- gineering 5 (2), 1-7.

[70] Vendite, C. C.; Vendite, L. L.; Moraes, A. C. de. 2005. Scout No Futebol: Uma Ferramenta Para a Imprensa Esportiva ?In:?, CONGRESSO BRASILEIRO DE CIENCIAS DA COMUNICAÇÃO. , 2005, Rio de Janeiro, UERJ. p. 1-10.

[71] Vendite, L. L.; Arruda M. 2012. Futebol: Ciências Aplicadas ao Jogo e ao Treinamento. 1. ed. São Paulo: Editiora Phorte Ltda. 560p.

[72] Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems 2 (1-3), 37-52.