UNIVERSIDADE FEDERAL RURAL DO SEMIÁRIDO - UFERSA
CURSO DE CIÊNCIA DA COMPUTAÇÃO
Trabalho de Conclusão de Curso (2021).

UTILIZAÇÃO DE TÉCNICAS DE DESCOBERTA DE CONHECIMENTO PARA TRAÇAR O PERFIL DE EQUIPES VENCEDORAS EM TORNEIOS DE ROBÓTICA.

Máspoly Gênes de M. Paiva *maspoly@ufersa.edu.br*
Angélica Félix de Castro *angelica@ufersa.edu.br*
Paulo Gabriel Gadelha Queiroz *pgabriel@ufersa.edu.br*

Resumo: As competições de robótica servem para que os alunos coloquem à prova os conhecimentos adquiridos durante o ano. Servem, também, como recurso de divulgação dentro e fora da escola, principalmente quando esta obtém resultados positivos, fazendo com que novos alunos ingressem na escola e se interessem pela robótica educacional. Este artigo trata sobre o uso de processo de descoberta de conhecimento em base de dados ou *Knowledge Discovery in Databases* – KDD utilizado em dados obtidos por meio de questionário com equipes diversas, campeãs e com desempenho considerados bons, para que, com a descoberta, fosse possível traçar o perfil de equipes vencedoras em torneios de robótica. Neste trabalho são comparados os algoritmos Apriori e *MeanShift*. Após alguns testes e associações, usando ambos os algoritmos, foi traçado um perfil que possibilita a criação de equipes vencedoras em competição de robótica. Os resultados obtidos demonstraram que a abordagem baseada no Apriori obteve a melhor performance para a formação da equipe de robótica.

Palavras-chave: Robótica; Descoberta de Conhecimento; Mineração de Dados; Regras de Associação.

1. INTRODUÇÃO

A robótica educacional, ou pedagógica, segundo Silva [1] "envolve um processo de motivação, colaboração e reconstrução" o que se faz necessário a utilização de conceitos que levam os alunos a um aprendizado interdisciplinar. A implantação da robótica educacional vem crescendo nos últimos anos e, como consequência, observa-se um aumento no número de eventos competitivos que buscam incentivar os alunos a desenvolver seus conhecimentos.

A robótica educacional, ao trabalhar conteúdos interdisciplinares com os alunos, pode, também, ser utilizada como ferramenta de ensino-aprendizagem na introdução de conceitos de engenharia e programação, de maneira lúdica e de uma forma amigável. Sua implantação tem alcançado grande aceitação desde as séries iniciais do ensino fundamental até a terceira série do ensino médio.

As competições são mais um ingrediente para chamar a atenção dos jovens a ingressar no mundo da robótica, pois na escola vencedora a robótica se populariza, fazendo surgir novos talentos. No Brasil duas competições se destacam: a Olimpíada Brasileira de Robótica – OBR e a FIRST® LEGO® *League* – FLL. Ambas competições têm fases regionais e nacionais, a FLL tem, ainda, uma fase internacional realizada com os melhores de cada país participante. Estas competições são geralmente realizadas com formações de equipes. Mas, com tantas competições, tantos alunos para participação, surge uma questão: como montar uma equipe com maior potencial para ser campeã?

É possível selecionar dentre os alunos/participantes do curso de robótica uma equipe com maior potencial para ser campeã? Quais critérios devem ser levados em consideração na hora de escolher os membros da equipe? Que fatores influenciam em um bom desempenho da equipe? As técnicas de descoberta de conhecimento em base de dados, ou *Knowledge Discovery in Databases* – KDD podem auxiliar na resolução destas questões. A técnica do KDD identifica padrões através de mineração de dados, descobrindo informações importantes, alcançando assim o objetivo do trabalho que é a formação de possíveis equipes vencedoras em campeonatos de robótica.

Uma nova geração de teorias e ferramentas computacionais estão surgindo para ajudar os seres humanos a extrair informações úteis [2]. Essas teorias e ferramentas são o tema do campo da descoberta de conhecimento de banco de dados (*Knowledge Discovery in Databases* – KDD) que se preocupa com o desenvolvimento de métodos e técnicas para dar sentido aos dados.

Segundo Cardoso e Machado [3], KDD compreende a todas as etapas de descoberta de conhecimento a partir de um banco de dados, ou seja, o conhecimento são as relações e padrões entre os elementos dos conjuntos de dados, diferente do termo *Data Mining* (Mineração de Dados) que é apenas uma das etapas do processo.

Esse trabalho tem como objetivo principal analisar, comparar e implementar dois algoritmos do KDD - os algoritmos Apriori e o *MeanShift*, o primeiro através da associação de dados e o segundo por clusterização - no problema da robótica, como forma de avaliar algoritmo é o mais adequado para gerar uma possível equipe campeã.

Este artigo está dividido da seguinte forma: na Seção 2 são apresentados alguns trabalhos relacionados; na Seção 3 está apresentada uma fundamentação teórica sobre assuntos importantes para compreensão desse trabalho.

Na Seção 4, os métodos e procedimentos utilizados para obtenção dos dados estão descritos, bem como a implementação dos algoritmos escolhidos. Na Seção 5 são exibidos os resultados obtidos e, por fim, na Seção 6, a conclusão e trabalhos futuros são apresentados.

2. TRABALHOS RELACIONADOS

A comparação entre algoritmos é realizada com frequência, principalmente em trabalhos nos quais se busca a retirada de conhecimento de um banco de dados. No trabalho de Neto et. al (2019) [4] busca-se realizar uma análise comparativa em monitoramento de integridade estrutural. Três abordagens de detecção de danos baseadas nos algoritmos *KMeans*, DBSCAN e *MeanShift* são comparadas. Os resultados gerados foram que apenas um dos algoritmos, o DBSCAN, satisfez e resolveu o problema.

Prass (2004) [5], assim como Neto et. al [4], faz diferentes comparações entre algoritmos para identificar qual é o mais eficaz para resolução de problemas no processo de descoberta de conhecimento ou KDD. Prass [5] faz paralelos entre algoritmos utilizando a mesma linguagem de programação e entre diferentes algoritmos com abordagem de resolução do problema.

Não são poucos os trabalhos em que os algoritmos Apriori e *MeanShift* são utilizados para classificação de grupos, uns por associação de características (Apriori) e outros por agrupamento de características (*MeanShift*). O trabalho de Maciel (2019) [6] busca traça o perfil de pessoas que cometem suicídios, classificando padrões através de regras de associação geradas pelo algoritmo Apriori. A Mineração de Dados foi a principal fase do KDD, utilizando técnicas diferentes para classificação dos dados [6] e então chegar ao objetivo do trabalho.

Lima (2015) [7] propõe um método alternativo para detecção de Ilhas Genômicas - IGs em bactérias do gênero Mycoplasmas, utilizando o método de clusterização *MeanShift*, fundamentando-se na hipótese de que é possível separar sequencias adquiridas das originais através do algoritmo de clusterização [7].

A utilização de *Data Mining*, etapa bastante importe para o KDD, pode ser vista em Cardoso e Machado (2007) [8] que empregou técnicas e funções específicas criadas para descobrir padrões de comportamento mais relevantes nos dados obtidos.

A temática da robótica educacional vem sendo tratada recentemente em vários trabalhos. Silva [1] trata sobre uma metodologia de ensino empregando robótica educacional aos alunos de ensino fundamental, que são público alvo das competições de robótica a nível nacional. Este trabalho também tem como temática a robótica, mas voltando-se para os campeonatos, uma vez que trata sobre como destacar alunos para formação de uma possível equipe campeã.

3. FUNDAMENTAÇÃO TEÓRICA

Existem diversas formas de trabalhar com dados, independentemente do tipo de dado. Uma dessas formas, que é utilizada nesta pesquisa, é a classificação ou formação de grupos (*clusters*) em determinadas categorias. Essa forma de análise dos dados facilita seu entendimento por conta de sua clareza e visualização [4].

A análise de dados de agrupamento pode ser dividida em duas partes [9]: análises exploratórias ou confirmatórias, usando modelos apropriados para a fonte de dados. Ambas as abordagens levam em consideração a classificação dos dados realizada a partir de uma das seguintes formas [9]:

- Ajuste a um modelo postulado, por exemplo, ao utilizar o algoritmo Apriori são identificados padrões, estes padrões podem ser usados como modelo para a análise de outros dados;
- Grupos naturais descobertos por meio de análise, um exemplo é mostrado na Figura 1. Cada cor (verde, amarelo e roxo) representa um grupo diferente, assim como cada ponto do plano pertence a um grupo específico, correspondendo à sua cor. Um grupo ou cluster pode ser resumido por uma média dos elementos pertencentes a esse grupo, comumente chamado de centroide de grupo [4].

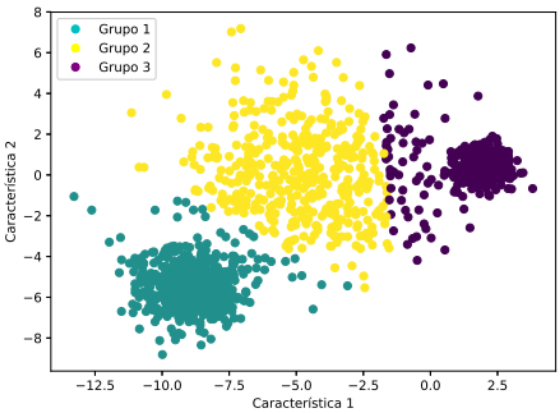

Figura 1 – Exemplo de Clusterização [10]

Cada algoritmo de agrupamento aplica sua técnica de classificação de acordo com o problema proposto, mas todas essas técnicas seguem uma similaridade no momento da classificação. Tal similaridade é melhor ilustrada com o algoritmo usado neste trabalho, o *MeanShift* [4].

A seguir são apresentados os princípios de funcionamento dos algoritmos de associação e de agrupamento assim como a abordagem para detecção de anomalias.

A. Regras de Associação e Algoritmo Apriori:

Muitos conjuntos de dados são retirados após o processo de *Data Mining*, para tal foi necessário o uso de tarefas de busca de informação. Uma destas tarefas são as Regras de Associação. Tais regras tem como tarefa identificar associações entre registros que são ou deveriam estar relacionados de alguma forma. Sua premissa básica é encontrar elementos que impliquem a presença de outros na mesma transação [11].

O requisito básico da tarefa de associação é encontrar elementos que impliquem a presença de outros elementos na mesma transação, ou seja, encontrar relacionamentos ou padrões comuns entre os registros. O termo transação indica quais itens foram consultados em uma operação de consulta específica.

Regras de associação geralmente representam padrões que existem em transações armazenadas. Por exemplo, uma estratégia de mineração que usa regras de associação poderia gerar a seguinte regra de um banco de dados de clientes que compraram itens: {leite, manteiga} → {pão}, que especifica que o cliente que compra leite e manteiga também compra pão com alguma certeza. Este nível de segurança para uma regra é definido por dois índices: o fator de suporte e o fator de confiança [11].

O algoritmo Apriori é um dos algoritmos mais conhecidos para mineração utilizando regras de associação [11]. O algoritmo usa uma pesquisa aprofundada e gera conjuntos de elementos candidatos (padrões) a partir de conjuntos de elementos k-1. Padrões raros são excluídos. Todo o banco de dados é pesquisado e conjuntos de itens comuns são recuperados dos conjuntos de itens candidatos [12].

B. Regras de Clusterização e Algoritmo MeanShift:

A clusterização – agrupamento – é uma tarefa que visa segmentar populações heterogêneas em subgrupos ou segmentos homogêneos. A formação de grupos de objetos em classes de objetos semelhantes é chamada de agrupamento, ou clusterização. Um grupo é uma coleção de objetos de dados que são semelhantes dentro do mesmo grupo [13].

O *MeanShift* é um algoritmo que utiliza as regras de agrupamento para minerar os dados e obter informações. É baseado nos centróides que funcionam atualizando os centróides candidatos de forma que os centróides permaneçam a média das observações dentro de uma determinada partição ou região. Os candidatos são filtrados para formar o conjunto final de prioridades [14].

4. MÉTODOS E PROCEDIMENTOS

Este trabalho foi desenvolvido utilizando o método de pesquisa descritivo-exploratório [15], uma vez que foi preciso o uso de dados coletados de participantes dos campeonatos de robótica e de referenciais teóricos para a contextualização sobre o que foi feito. Ainda segundo Moretti [15], a pesquisa pode ser classificada quanto ao seu tipo de abordagem como quali-quanti, pois seus resultados são expressos tanto utilizando técnicas estatísticas, quantitativas, como qualitativas, pois também apresenta resultados através de percepções e análises.

Foram coletados dados de membros de 12 (doze) equipes que participaram de campeonatos de robótica, dentre estes, 4 (quatro) equipes campeãs. Estes dados foram utilizados em algoritmo de associação para assim descobrir, por meio de algoritmos de descoberta de conhecimento em banco de dados (KDD), quais são as principais características que a fizeram campeã.

Os dados coletados passaram pelas etapas do KDD desta forma, primeiramente pela etapa de seleções na qual foi decidido quais dados são relevantes para a formação de resultados com informações úteis. Após esta etapa, os dados foram para o pré-processamento, etapa em que houve limpeza de dados e seleção de atributos, ou seja, dados inconsistentes ou errôneos foram corrigidos para não comprometer a qualidade do modelo de conhecimento extraído pelo KDD e desta forma passa pelas etapas de transformação e mineração de dados sendo reorganizados de uma forma específica. Primeiramente foram aplicados os algoritmos Apriori e *MeanShift* nos dados das equipes campeãs para traçar o perfil dos alunos. Após traçar o perfil, foi utilizado novamente os algoritmos no restante dos dados para identificar alunos para formar uma possível equipe campeã. Esse perfil foi formado lendo as associações geradas pelo Apriori e pelo *MeanShift*, chegando na etapa final que é a interpretação de resultados onde surgem padrões de relacionamentos e descobertas de novos fatos [16] [17].

Na etapa de Mineração de Dados há aplicação dos algoritmos escolhidos neste trabalho. O Apriori com suas regras de associação é utilizado, com os padrões conhecidos, e há uma avaliação das informações, por fim o conhecimento extraído. De igual forma, os dados são tratados com *MeanShift*, desta vez as regras de clusterização são utilizadas, segundo o algoritmo implementado, novamente há um tratamento das informações e então produzido o conhecimento.

As etapas utilizadas para gerar conhecimento são explicadas com mais detalhes nas próximas subseções: 4.1 – que trata da coleta de dados; 4.2 – que mostra como cada etapa do KDD foi executada e seus resultados.

4.1. Coleta de Dados

Os dados utilizados foram coletados por meio de questionários respondidos pelas equipes participantes de competições regionais e nacionais da Olimpíada Brasileira de Robótica – OBR e da FIRST LEGO *League* – FLL. O questionário possui perguntas estruturadas para facilitar as inserções dos dados no banco de dados.

O questionário incluía desde questões socioeconômicas até questões que envolvem a vida estudantil, uma vez que os membros das equipes são alunos de escolas tanto públicas quanto particulares. O questionário pode ser acessado através do link: <https://forms.gle/4PrqU36ynnM8FM5G7>.

4.2. *Knowledge Discovery in Databases* – KDD

O KDD é dividido em cinco etapas com objetivo de descobrir informações consideráveis que apoiem os tomadores de decisão em suas estratégias, Figura 2.

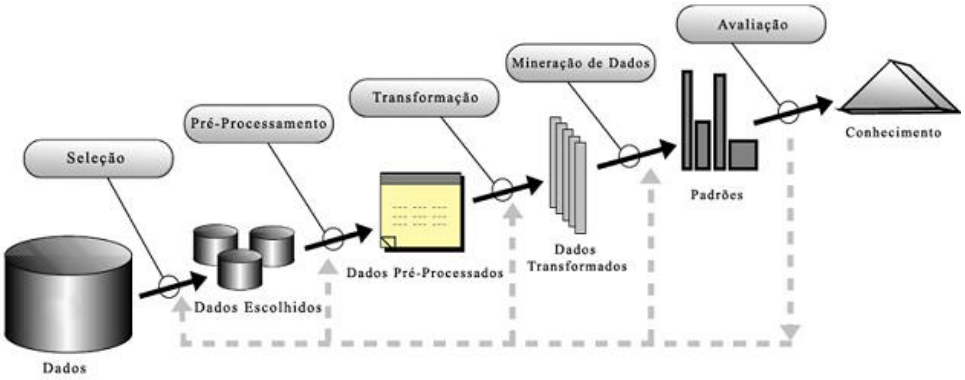

Figura 2 – Etapas do KDD [18]

4.2.1. Seleção

Etapa na qual, segundo Lira [18], as fontes relacionadas ao domínio são definidas para extrair os dados apropriados para o contexto.

Nesta etapa foram decididos quais dados são relevantes para que sejam obtidos resultados com informações úteis. Em nossa pesquisa foram obtidas 25 (vinte e cinco) respostas que correspondem ao mesmo número de linhas em uma tabela em Linguagem de Consulta Estruturada (*Strutured Query Language* – SQL), e 31 (trinta e uma) colunas que são correspondentes ao número de perguntas do questionário mais uma gerada automaticamente, totalizando 32 (trinta e duas) colunas.

A seleção aconteceu da seguinte maneira: primeiramente foram excluídas colunas de identificação, "Carimbo de data/hora" que foi gerada automaticamente e a coluna "Endereço de e-mail". Essas colunas não serviriam para processo de KDD uma vez que são únicas e desta forma não gerariam regras de associação. Após a exclusão viu-se que as colunas: "idade", "sexo", "cidade_reside", "estado", "área", "com_quem_mora", "situacao_pais", "renda_familiar", "tem_na_casa", "internet", "escola_atual", "estudou_noutra_escola", "escola_publica", "bom_aluno", "recuperação", "qual_disciplina", "português", "matematica", "historia", "geografia", "ingles", "ciencias", "estuda_robotica", "horas_pratica" e "equipe" são necessárias para o processo KDD, Figura 3.

idade	sexo	cidade_reside	estado	area	com_quem_mora	situacao_pais	renda_familiar	tem_na_casa	internet
escola_atual	estudou_noutra_escola	escola_publica	bom_aluno	recuperacao	qual_disciplina	portugues	matematica	historia	geografia	ingles	ciencias	estuda_robotica	horas_pratica	equipe

Figura 3. Colunas utilizadas após Fase Seleção. (Autoria própria)

4.2.2. Pré-processamento

O subconjunto de dados selecionado pode ter alguns erros, como: dados ausentes, dados inválidos, registros duplicados e ruído [18].

Nesta etapa alguns dados foram corrigidos. No campo "qual_disciplina" foram feitas algumas correções para que pudesse ser utilizado para gerar conhecimento. Respostas do tipo "Várias kkk" e "Faz pergunta difícil não, kkkk" foram corrigidas para "Várias". Outras correções de digitação ou acentuação foram feitas em campos diversos.

Outros campos também foram corrigidos, como “equipe”, “material_usa” e “ano_participacao”, foram retiradas informações repetidas ou sem necessidade para o KDD. O campo “equipe” foi corrigido da seguinte forma: a equipe do entrevistado ficou somente aquela mais recente, por exemplo: o entrevistado participou das equipes “Legowork (2018) e Monarcas Absolutistas (2019)”, então foi utilizada a equipe mais recente, de maneira que este entrevistado ficou com o campo equipe “Monarcas Absolutistas”, na Figura 4 temos o antes e depois da correção.

equipe
Legowork (2018) e Monarcas Absolutistas (2019)

→

equipe
Monarcas Absolutistas

Figura 4. Coluna “equipe” antes e depois do pré-processamento. (Autoria própria)

4.2.3. Transformação

Fase em que os dados devem ser formatados e armazenados corretamente [18]. O principal objetivo é resolver problemas de unidade, escala e padronização de formatos [6].

Na etapa de transformação alguns campos foram adequados para aplicação das regras de associação. O campo “Na sua casa tem” foi dividido em 6 (seis) novos campos, “tv”, “internet”, “computador”, “notebook”, “smartphone”, “tablet”, no qual as respostas passaram a ser como mostrado na Figura 5.

tv	internet	computador	notebook	smartphone	tablet
TV	NET	COMP	NNOT	NFON	TAB
TV	NET	NPC	NOTE	FONE	NTAB
TV	NET	COMP	NOTE	FONE	NTAB

Figura 5. Colunas após transformação. (Autoria própria)

O campo “Smartphone com Internet” foi excluído devido a redundância de informação, pois após a transformação do campo “Na sua casa tem” em 6 novos campos, percebeu-se que aqueles possuíam smartphone também possuíam internet.

Para o campo “idade” foi criada uma função que retorna a faixa etária dos competidores, visto que as competições possuem níveis e estes são divididos de acordo com a faixa etária.

4.2.4. Mineração de Dados

A mineração de dados é a fase principal do KDD. Quando ocorre efetivamente a busca de conhecimento novo e útil a partir dos dados e busca conhecimento abstrato. Para cada tipo de conhecimento que se deseja obter, existe um algoritmo diferente que pode ajudar a atingir o objetivo final, por exemplo: associação e clusterização [6].

A etapa de mineração de dados consiste em aplicar algoritmos de análise e descoberta de dados que produzem vários modelos [3], de maneira que estes modelos possam produzir conhecimento. Neste trabalho, os algoritmos Apriori e *MeanShift* foram implementados no estudo de caso.

Na etapa de mineração de dados, os dados colhidos na pesquisa, selecionados, pré-processados e transformados passaram pelo algoritmo de associação de dados para que pudesse ser extraído informações e gerar conhecimento. O mesmo processo foi repetido para o algoritmo de clusterização. Para Vasconcelos e Carvalho [11], a mineração de dados é “o núcleo de processo de descoberta de conhecimento em banco de dados”.

4.2.4.1 Aplicação do Algoritmo Apriori

O Apriori foi um dos algoritmos selecionados para uso neste trabalho por ser de fácil implementação e pela capacidade de trabalhar com associação em grandes bancos de dados, encontrando conjuntos de itens frequentes mantendo um bom desempenho [19].

O algoritmo Apriori usa dois índices que auxiliam na associação, são eles: suporte e confiança.

Suporte é a proporção de transações de um ou mais itens juntos sobre o número total de transações, conforme Fórmula 1.

$$\text{Suporte} = \text{Nº de registros da tabela que contêm o dado} \,/\, \text{Nº total de registros da tabela}$$

Fórmula 1: Definição de Suporte [20]

Por exemplo, na base de dados deste trabalho podemos pegar o suporte do item {bom_aluno} dentre os entrevistados que fazem parte de equipe campeã. Para calcular o suporte tem-se que ver quantas vezes o valor “Sim” aparece nesta coluna e dividir pelo total de registros observados. Observa-se esses valores na Tabela 1.

faixa_etaria	sexo	estudou_noutra_escola	bom_aluno	recuperacao
De 14 - 16 anos	Masculino	Sim	Sim	Não
De 14 - 16 anos	Feminino	Sim	Sim	Não
De 17 - 18 anos	Masculino	Não	Sim	Não
De 14 - 16 anos	Masculino	Sim	Sim	Não
De 14 - 16 anos	Masculino	Não	Não	Sim
De 17 - 18 anos	Masculino	Não	Não	Sim
De 11 - 13 anos	Masculino	Sim	Sim	Não
De 14 - 16 anos	Masculino	Sim	Sim	Não
De 14 - 16 anos	Masculino	Sim	Sim	Sim
De 11 - 13 anos	Masculino	Sim	Sim	Não

Tabela 1 – Associações dos dados sociais das equipes campeãs pesquisadas (Autoria Própria)

Na Fórmula 2, podemos observar o cálculo realizado. São 8 (oito) respostas "Sim" na coluna {bom_aluno} de um total 10 (dez) registros observados. O que nos dá um suporte de 80%. Para uso no algoritmo Apriori, determinamos o Suporte como valor mínimo.

$$\text{Suporte} = \frac{8}{10} \rightarrow \text{Suporte} = 0{,}8 = 80\%$$

Fórmula 2: Obtenção do Suporte para respostas "Sim" em {bom_aluno}. (Autoria Própria)

A confiança é, de maneira semelhante, a porcentagem de associações que contém a resposta A e B mas não necessariamente a reposta B. O valor utilizado para formar as regras são determinados pelo usuário, o algoritmo realiza os cálculos e então faz as associações, Fórmula 3.

$$\text{Confiança} = \frac{\text{Número de registros da tabela que contêm todos os itens da regra}}{\text{Nº de registros da tabela que contêm o antecedente da regra}}$$

Fórmula 3: Definição de Confiança [20]

Seguindo o exemplo com nossos dados, confiança é a porcentagem da resposta "Sim" em {recuperação} se a resposta for "Sim" em {bom_aluno} e da mesma forma de maneira inversa, quantos "Sim" em {bom_aluno} se a resposta for "Sim" em {recuperação}, Fórmula 4. O algoritmo também utiliza a confiança mínima para gerar as melhores regras.

$$\text{SE sim \{bom\_aluno\} ENTÃO sim \{recuperação\}: } \text{Confiança} = \frac{1}{8} \rightarrow \text{Confiança} = 0{,}125 = 12{,}5\%$$

$$\text{SE sim \{recuperacao\} ENTÃO sim \{bom\_aluno\}: } \text{Confiança} = \frac{1}{3} \rightarrow \text{Confiança} = 0{,}333 = 33{,}3\%$$

Fórmula 4: Cálculo da Confiança (Autoria Própria)

Neste caso poderia se utilizar uma confiança mínima de 33% para determinar as outras associações, claro que analisando todo a base de dados pode-se admitir uma confiança maior como limite mínimo, o que deixaria de fora associações com baixo índice de confiança.

Há também um novo limite chamado de *Lift*, que é usado em conjunto com o suporte mínimo e a confiança

mínima que tem a fórmula apresentada em Fórmula 5 [20]. É utilizada para ordenar as combinações. O valor do *Lift* expressa a relação de quanto maior o número melhor.

$$Lift = \frac{\text{Confiança da regra}}{\text{Suporte do consequente da regra}}$$

Fórmula 5: Definição de *Lift* [20]

Continuando o exemplo com a nossa base de dados, determinamos o *Lift* dividindo a confiança pelo suporte, onde confiança é maior que suporte. Na Tabela 2 podemos observar as confianças e os suporte obtidos da base de dados apresentados na Tabela 1.

Suporte			Confiança		*Lift*		
faixa_etaria	De 14 - 16 anos	60%	SE sexo(masculino)	Então estudou_noutra_escola (sim)	confianca - SE sexo(masculino) ENTÃO estudou_noutra_escola (sim)	suporte estudou_noutra_escola (sim)	total
	De 17 - 18 anos	20%	86%		0,86	0,7	1,23
	De 11 - 13 anos	20%	SE bom_aluno(sim)	ENTÃO recuperação (não)	confianca - SE bom_aluno (sim) ENTÃO recuperação (não)	suporte recuperação (não)	total
sexo	masculino	90%	88%		0,88	0,7	1,26
	feminino	10%	Suporte = Quantidade de vezes que o item aparece / Quantidade total de respostas				
estudou_noutra_escola	sim	70%					
	não	30%	Confiança = Quantidade de vezes em que os itens aparecem juntos / Quantidade de vezes que o item do SE aparece				
bom_aluno	sim	90%					
	não	10%					
recuperacao	sim	30%	*Lift* = Confiança (A>B) / Suporte do Item do ENTÃO				
	não	70%					

Tabela 2 – Parte dos resultados do suporte, da confiança e do *lift* dos dados sociais das equipes campeãs pesquisadas (Autoria Própria)

4.2.4.2 Aplicação do Algoritmo Mean Shift

O *MeanShift* foi escolhido para utilização neste trabalho por ser um algoritmo de clusterização e por ser, assim como Apriori, de fácil implementação. A clusterização tem por objetivo identificar e agrupar registros semelhantes em um banco de dados, formando *clusters* (grupos) [9]. É vantajoso utilizar o *MeanShift* pois ele escolhe automaticamente os números e a forma dos *clusters* e possui um bom desempenho.

Não há a necessidade de o usuário definir parâmetros, exceto os de resolução da análise [21]. Parâmetros como largura de banda – *bandwidth* e *seeds* são calculadas internamente na classe *Meanshift*. Se o *bandwidth* não for fornecido, a largura de banda é estimada utilizando uma função da classe *Sklearn*. Todas essas facilidades na implementação, fez com o algoritmo do *MeanShift* fosse escolhido para ser utilizado no trabalho.

4.2.5. Avaliação

Através desta fase é possível chegar à informação desejada, interpretando e avaliando os padrões encontrados pela fase Mineração Dados. Os usuários podem usar várias ferramentas com funções estatísticas e de visualização para validar ou avaliar um padrão irrelevante [18].

Detalhes sobre esta etapa são apresentadas na Seção 5 – Resultados e Discussões.

5. RESULTADOS E DISCUSSÕES

Foram obtidos dados através dos algoritmos Apriori e *MeanShift*. Nas subseções 5.1 – Algoritmo Apriori e 5.2 – *MeanShift* são apresentados os resultados correspondentes.

5.1. Apriori

Foram obtidos dados através do algoritmo Apriori utilizando índices de suporte mínimo de 15% (quinze porcento), ou seja, valor mínimo de suporte que deve ser utilizado para gerar regras de associação e confiança de no mínimo de 80% (oitenta porcento), assim como o suporte mínimo, somente valores de no mínimo 80% de confiança são considerados relevantes para criação das regras de associação. Houve mudança nos índices do *Lift* conforme os dados analisados. Com os valores determinados de suporte e confiança e com *Lift* mínimo de 5,0 (cinco) não foi possível obter associação. Porém com *Lift* mínimo de 2,0 (dois) foi possível extrair, dos dados sociais, 2 (duas) associações, que foram analisadas e pode-se concluir que a questão social, para estes dados pesquisados não tiveram grande influência nos resultados. Isto deve-se ao fato de os entrevistados serem todos alunos de escolas particulares. Na Tabela 3 podemos observar as associações geradas.

SE	ENTÃO	Confiança	*Lift*
{'Acima de R$9.001,00', 'Casados'}	{'FONESim', 'Masculino', 'NETSim', 'Pais'}	1.0	2.08
{'Casados', 'De 14 - 16 anos', 'Entre R$2.601 até R$ 4.000,00'}	{'FONESim', 'Masculino', 'NETSim', 'Pais'}	1.0	2.08

Tabela 3 – Associações dos dados sociais das equipes campeãs pesquisadas (Autoria Própria)

O algoritmo executado com os dados sociais também foi utilizado com os dados escolares onde são questionados os desempenhos dos alunos em sala de aula assim como perguntas relacionadas ao estudo de robótica e suas práticas. Utilizando os mesmos parâmetros foi possível observar os seguintes resultados conforme Tabela 4.

Parâmetros	Associações Obtidas					
	Dados Gerais	Campeãs dados Gerais	Dados Escolares	Campeãs dados Escolares	Dados Sociais	Campeãs dados Sociais
c = 80%, s = 15%, l = 2,0	1961	28160	34	720	2	28
c = 80%, s = 15%, l = 5,0	28	5792	0	64	0	0

Tabela 4 – Associações dos dados pesquisados (Autoria Própria)

Na Tabela 4, observamos o total de associações obtidas tanto com os dados das equipes campeãs, como com os dados de todos os pesquisados. O algoritmo realizou 3 (três) busca por associações com cada grupo, campeões e todos pesquisados, sendo uma vez com os dados gerais – todos os dados obtidos em cada grupo, outra vez com apenas os dados socioeconômicos e mais uma vez somente com os dados escolares. Os valores de confiança e suporte mantiveram-se os mesmos – c= 80% e s=15%, alterando apenas o *lift* – l=2,0 e l=5,0.

No Gráfico 1, podemos observar os dados em comparação e nota-se a diferença da quantidade de associações mudando o valor do *Lift* mínimo, uma vez que este parâmetro quanto maior, maior também será a relevância das associações.

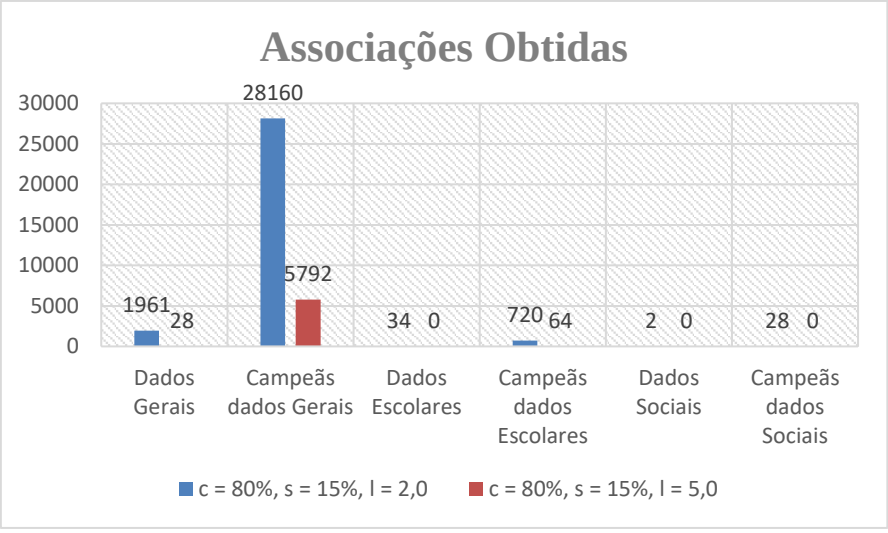

Gráfico 1 – Associações Obtidas (Autoria Própria)

No Gráfico 1 temos c=80% que indica uma confiança mínima de 80% (oitenta porcento) das associações bem como s=15% indica suporte mínimo de 15% (quinze porcento) e l=2,0 e l=5,0 indicam que para obtenção das associações foram consideradas as regras que tinham no mínimo de *lift* de 2,0 e *lift* de 5,0 respectivamente.

Ao analisar os dados escolares das equipes campeãs observou-se que os integrantes são bons alunos, preferem disciplinas de exatas e idiomas, já estudam robótica entre 2 e 3 anos e que o Kit LEGO EV3 é o mais utilizado. Também pode-se perceber que os campeões treinam mais robótica, mais de 3 horas semanais. Na Tabela 5 um exemplo de associação gerada. Pode-se notar, também, a presença massiva do gênero "masculino" o que espelha a realidade de que é a baixa participação das meninas na robótica, dados da própria organização da OBR demonstra que no ano de 2020, na modalidade prática, as meninas eram apenas 30% dos participantes [22].

SE	ENTÃO	Confiança	*Lift*
{'2P', '3I', '4G', 'Masculino', 'RNão', 'BSim'}	{'5M', 'Entre 2 e 3 anos', 'Kit LEGO EV3'}	1.0	5.0
{'2P', '3I', '4H', 'Kit LEGO EV3', 'Masculino', 'RNão'}	{'5M', 'Entre 2 e 3 anos', 'BSim'}	1.0	5.0
{'3P', 'Mais de 3 horas', 'RNão', 'BSim'}	{'4I', 'Kit LEGO EV3'}	1.0	2.5
{'3I', 'Kit LEGO EV3', 'Mais de 3 horas', 'RNão'}	{'4M', 'BSim'}	1.0	2.5

Tabela 5 - Associações dos dados escolares das equipes campeãs pesquisadas (Autoria Própria)

As numerações '2P', '3P', '3I', '4G', '4H' da coluna "SE" referem-se as preferências das disciplinas de português, inglês e história/geografia, o 'RNão' indica se o aluno ficou alguma vez em recuperação já na coluna "ENTÃO" o 'BSim' indica que o aluno se considera bom aluno, o '4I', '4M', '5M' foram as notas de preferência em inglês e matemática respectivamente.

Os dados analisados em conjunto, dados sociais e escolares, resultou em 28160 (vinte e oito mil cento e sessenta) associações, ou seja, uma quantidade expressiva. Dessa forma foram alterados os índices de suporte para 20%, confiança para 80% e *Lift* para 20%, pode-se obter um número de associações que se aproxima da realidade. As associações apresentadas foram no total de 136 (cento e vinte quatro), apenas 1% (um porcento) do total de associações geradas com parâmetros diferentes, que foram analisadas para retirada de informação. Pôde-se observar que os integrantes de equipes campeãs possuem entre 14 e 16 anos, os pais são casados, possuem internet, televisão, *notebook/laptop*, estudam robótica pelo menos uma hora por dia, já estudam robótica há 3 anos, são bons aluno e não ficaram em recuperação.

Com as informações coletadas foi possível construir o conhecimento de como montar uma equipe campeã de robótica. Para ser um possível campeão em robótica os estudos de robótica devem começar cedo, por volta dos dez anos de idade, mas não basta começar cedo, deve-se manter um ritmo de estudos diários de pelo menos uma hora. Bons alunos têm mais chances de se tornarem campeões. Uma família estruturada também auxilia no desempenho dos alunos. A conexão com internet é necessária, seja ela feita através de *smartphone* e/ou por meio de *notebooks/laptops*, usada como ferramenta facilitadora de aprendizagem e que horas extras de treinamento de robótica ajuda no desempenho da equipe durante as competições.

5.2. *MeanShift*

Os resultados obtidos com o algoritmo *MeanShift* não foram satisfatórios. Todo o processo de mineração realizado foi semelhante ao que foi executado com o algoritmo Apriori.

Com o *MeanShift* o resultado foi de apenas 3 (três) grupos, não importava o dado que se escolhia como parâmetro, sempre foram os mesmos 3 grupos, um exemplo de resultado é que quando se utilizou todos os alunos – 25 (vinte e cinco) – um grupo ficou composto por 23 (vinte e três) participantes e os outros 2 (dois) grupos compostos por 1(um) aluno cada. A configuração dos grupos não alterava por mais que os parâmetros mudassem.

6. CONCLUSÕES

A partir do exposto, pode-se perceber que é possível traçar o perfil de uma equipe campeã de robótica. A utilização KDD, é muito importante pois serviu para nortear o processo de descoberta.

Os resultados colhidos são de grande valia, com eles pode-se forma e preparar as equipes para que possam ser campeãs. Uma equipe não é formada apenas pelos membros que a compõem, ou pelos anos que pratica/estuda robótica, é formada também pela estrutura familiar e pelo empenho dos alunos nas disciplinas curriculares da escola.

O algoritmo Apriori demonstrou-se uma excelente ferramenta de associação de dados. Com as associações geradas foi possível gerar conhecimento. Infelizmente não foi possível comparar com o algoritmo de clusterização, *MeanShift* pois este não gerou resultados expressivos.

Os resultados gerados com o Apriori, por outro lado, são de grande importância e foi colocado em prova na Olimpíada Brasileira de Robótica – OBR quando uma equipe, que no ano de 2020 ficou na posição 16 na colocação geral, subiu para a 3ª colocação em 2021.

Os alunos em questão possuíam as características de equipe campeã a diferença entre um ano e outro foi o tempo de treinamento dedicado para a competição.

Ainda é possível melhorar o conhecimento com novos dados obtidos e assim tornar mais eficaz a utilização do KDD para formar possíveis equipes campeãs. Para trabalhos futuros incluir na pesquisa alunos de escolas da rede pública de ensino e mais itens no questionário para que desta forma fique mais precisa a caracterização dos membros de uma equipe campeã.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] SILVA, Alzira Ferreira da. RoboEduc: uma metodologia de aprendizado com robótica educacional. 2009. 133 f. Tese (Doutorado em Engenharia Elétrica /Área Engenharia da Computação) - Universidade Federal do Rio Grande do Norte, Natal, 2009. Disponível em: <ftp://ftp.ufrn.br/pub/biblioteca/ext/bdtd/AlziraFS.pdf>. Acesso em: 29 abril 2019.

[2] CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. Rev. Adm. Pública, Rio de Janeiro, v. 42, n. 3, p. 495-528, jun. 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122008000300004&lng=pt&nrm=iso>. Acesso em: 24 abril 2020.

[3] FAVERO, Luiz P. KDD e Data Mining: mais do que apenas conceitos. IT Fórum 365. 2019. Disponível em: <https://www.itforum365.com.br/colunas/kdd-e-data-mining-mais-do-que-apenas-conceitos/>. Acesso em: 20 abril 2020.

[4] NETO, R. M. B. et. al. Análise Comparativa da Aplicação de Técnicas de Clusterização no Monitoramento da Integridade Estrutural. CBIC 2019. Disponível em: <https://sbic.org.br/wp-content/uploads/2019/12/CBIC2019-55.pdf>. Acesso em: 05 setembro 2021.

[5] PRASS, Fernando Sarturi. Estudo Comparativo Entre Algoritmos de Análise de Agrupamentos em Data Mining. Dissertação (mestrado), Universidade Federal de Santa Catarina, 2004. Disponível em: < http://repositorio.ufsc.br/xmlui/handle/123456789/87267>. Acesso em: 15 setembro 2021.

[6] MACIEL, Laís R. B. Análise Comparativa Entre Algoritmo J48, Naivebayes e Apriori Para Base De Dados de Suicídio. Rev. UniScientiae, UNIVIÇOSA, v. 2, n. 2, p.137-153 jul./dez. 2019. Disponível em: < https://academico.univicosa.com.br/revista/index.php/RevistaTecnologiaeCiencia/article/view/1311/1458>. Acesso em: 28 outubro 2021.

[7] LIMA, Yuri Vinicius Verissimo de Lima. Predição e Caracterização de Ilhas Genômicas em Bactérias do Gênero Mycoplasmas, Utilizando o Método de Clusterização Mean Shift. 2015. 70f. Monografia (Trabalho de Conclusão de Curso). Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/5544>. Acesso em: 27 outubro 2021.

[8] CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. 2007. Disponível em: <https://www.scielo.br/j/rap/a/4ScBD9DkFprnH7MFyKC3ydv/?lang=pt>. Acesso em: 29 outubro 2021

[9] CAVALCANTI JUNIOR, Nicomendes L. Clusterização baseada em algoritmos fuzzy. Dissertação (mestrado). Universidade Federal de Pernambuco, p.21, 2006. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/2619/2/nlcj.pdf>. Acesso: 05 setembro 2021.

[10] Scikit learn: Clustering. 2019. Disponível em: <https://scikit-learn.org/stable/modules/clustering>. Acesso: 05 setembro 2021.

[11] VASCONCELOS, L. M. R.; CARVALHO, C. L. Aplicação de Regras de Associação para Mineração de Dados na Web. Relatório Técnico - Instituto de Informática da Universidade Federal de Goiás, Goiânia. 2004. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acesso em: 20 maio 2021.

[12] AGRAWAL, R; IMIELINSKI, T; SWAMI, A. Mining Association Rules between Sets of Items in Large

Databases. In: ACM SIGMOD CONFERENCE ON MANEGEMENT OF DATA, p. 207 – 216, Washington, DC, USA, 1993. ACM Press - New York, NY, USA.

[13] HAN, Jiawei ; KAMBER, Micheline; PEI, Jian. Data Mining: Concepts and Techniques. 3rd ed. The Morgan Kaufmann Series in Data Management Systems (Selected Titles). Elsevier, USA, August 2012.

[14] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 24, n. 5, pp. 603-619, May 2002, doi: 10.1109/34.1000236. Disponível em: <https://ieeexplore.ieee.org/document/1000236/references#references>. Acesso em: 31 outubro 2021.

[15] MORETTI, I. Metodologia de Pesquisa do TCC: conheça os tipos e veja como definir. Via Carreira. 2020. Disponível em: <https://viacarreira.com/metodologia-de-pesquisa-do-tcc/>. Acesso em: 16 abril 2020.

[16] DANTAS, E. R. G. et al. O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões. SEGeT – Simpósio de Excelência em Gestão e Tecnologia. Disponível em: <https://www.aedb.br/seget/arquivos/artigos08/331_331_Artigo_SEGET_EJDR_Versao_Final_010808.pdf>. Acesso em: 20 abril 2020.

[17] SCHEUNEMANN, F.; PRETTO, F. Mineração de dados para descoberta de conhecimento na Área de oncologia. Rev. Destaques Acadêmicos, Rio Grande do Sul, v. 7, n. 4, p. 51-67, 2015. Disponível em: <http://www.univates.br/revistas/index.php/destaques/article/download/499/491>. Acesso em: 24 abril 2020.

[18] LIRA, Kaynan C.; OLIVEIRA, Marcos A. de; MAGALHÃES, Regis P.; GONÇALVES, Enyo José T. Utilizando Mineração De Dados E Sistemas Multiagentes Na Análise Da Evasão Em Educação A Distância Por Meio Do Perfil Dos Alunos. Universidade Federal do Ceará (UFC). Disponível em: <https://www.researchgate.net/profile/Marcos-De-Oliveira-3/publication/308995146_Utilizando_Mineracao_de_Dados_e_Sistemas_Multiagentes_na_Analise_da_Evasao_em_Educacao_a_Distancia_por_meio_do_Perfil_dos_Alunos/links/57fd407808aeea8c97c87a8b/Utilizando-Mineracao-de-Dados-e-Sistemas-Multiagentes-na-Analise-da-Evasao-em-Educacao-a-Distancia-por-meio-do-Perfil-dos-Alunos.pdf>. Acesso em: 16 agosto 2021.

[19] ROMÃO, W et al. Extração De Regras De Associação Em C&T: O Algoritmo Apriori. Programa de Pós-Graduação em Engenharia de Produção – PPGEP. Disponível em: <http://www.abepro.org.br/biblioteca/ENEGEP1999_A0901.PDF>. Acesso em: 16 agosto 2021.

[20] SILVA, Glauco Carlos. Mineração de Regras de Associação Aplicada a Dados da Secretaria Municipal de Saúde de Londrina – PR. Dissertação (mestrado). Porto Alegre, 2004. f. 94, p.20. Disponível em: < https://www.lume.ufrgs.br/bitstream/handle/10183/8696/000586835.pdf?sequence=1>. Acesso em: 02 novembro 2021

[21] COMANICIU, Dorin; MEER, Peter. Mean Shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. pp. 603-619. Disponível em: <https://ieeexplore.ieee.org/document/1000236>. Acesso em: 01 novembro 2021.

[22] Mundo Robótica. A Robótica e a Pandemia: Novas Soluções Foram Desenvolvidas Rapidamente Para A Nova Realidade Mundial. Revista Oficial da Olimpíada Brasileira de Robótica, v. 07, n. 18, fev. 2021. São Carlos, SP. 2021.