**Autor/E-mail:** Igor Lima Rocha / [igor\_roc@hotmail.com.br](mailto:igor_roc@hotmail.com.br)

**Coordenador/E-mail:** Hélder Conceição Almeida / [helderca@uesc.br](mailto:helderca@uesc.br)

**Instituição:** Universidade Estadual de Santa Cruz

**Data:** 10/08/2019

**Local:** Ilhéus – BA

**Possibilidade de utilizar recursos de mineração de dados para prever perfis de evasão escolar**

**Resumo**

Este projeto busca mostrar a possibilidade de extrair e analisar dados sobre os estudantes de graduação na UESC, de forma a compreender as possíveis causas do índice de evasão do referido instituto. Neste projeto busca-se evidenciar que os conhecimentos podem ser utilizados e modelados como classificadores para identificar os alunos que estão mais propensos a abandonar a graduação. Serão usados três algoritmos: NaiveBayes, SupportVectorMachine e J48. Três abordagens de atributos também serão testadas: Manual, Seleção Baseada em Correlação e Ganho de Informação. Desse modo, a partir do entendimento do problema, tomar medidas na tentativa de reduzir a taxa de evasão, propondo auxílio ao aluno evasor antes do mesmo abandonar o curso.

**Palavras-chave:** NaiveBayes, SVM, J48, UESC, evasão, mineração de dados.

**Abstract**

This project seeks to show a possibility of extract and analise data about UESC degree students, in order to understand the possible causes of the dropout rate of the said institute. This project seeks to demonstrate that the knowledge can be used and modeled as classifiers to identify students who are more likely to drop graduation. Three algorithms also will be used: NaiveBayes, SupportVectorMachine and J48. Three attribute approaches will also be tested: Manual, Correlation Based Selection, and Information Gain. Thus, based on the understanding of the problem, take measures in an attempt to reduce the dropout rate, proposing assistance to the student evader before leaving the course.

**Key-words:** NaiveBayes, SVM, J48, UESC, dropout, data mining.

**1. INTRODUÇÃO**

A educação é um dos recursos mais importantes para que um país seja bem-sucedido, mas atualmente esse setor está sendo sucateado. Considerando que a evasão de alunos em cursos superiores tem se tornado um fenômeno de extrema necessidade de atenção Diante disso, fez-se necessária a busca por respostas para as possíveis causas de tal fenômeno.

O seguinte estudo busca analisar as possíveis implicações e limitações para a continuação e formação do aluno nos cursos de graduação na Instituição a qual está sendo utilizada como objeto de estudo (UESC), além da extração de dados úteis para a verificação e conclusão de possíveis agentes causadores da evasão. Assim, o problema central da pesquisa pode ser expresso com o questionamento: De que forma os algoritmos, como NaiveBayes, SupportVectorMachine e J48, podem auxiliar na análise do perfil de possíveis evasores dos cursos de graduação na UESC?

Ademais o problema de pesquisa central, outra questão específica também norteará o desenvolvimento desta pesquisa: Além de fatores sociais e econômicos, quais são os possíveis fatores que influenciam na construção do perfil de alunos que tendem a evasão? O objeto de pesquisa será constituído pelo estudo de tais perfis e levantamento de dados, de forma a criar um parâmetro o qual irá se encaixar no quadro extremo, facilitando a tomada de medidas auxiliares para evitar a possível evasão.

Decorrente do problema de pesquisa, algumas hipóteses nortearão o desenvolvimento da investigação aqui proposta. São elas:

* Evasão como descoberta de conhecimento.
* A mineração de dados como ferramenta para classificação de dados de alunos propensos à evasão.
* Perfil dos alunos: fatores que influenciam no abandono do curso.
* Classificação dos dados: problemas recorrentes e soluções viáveis a partir do uso da tecnologia computacional.
* Aplicações das técnicas em áreas mais amplas da educação.

**2. OBJETIVOS**

**2.1 Objetivo geral**

A pesquisa tem como objetivo geral analisar os dados armazenados na referida IE a fim de classificar, por meio de algoritmos, o perfil de alunos evasores e não evasores de forma mais precisa, assim possibilitando a tomada de contramedidas auxiliares aos potenciais evasores os quais estão matriculados na Universidade antes do abandono.

**2.2 Objetivos específicos**

São objetivos específicos deste projeto de pesquisa:

* Verificar os dados dos alunos contidos no banco de dados da Universidade para traçar perfis.
* Possibilitar a criação de um parâmetro de comparação para a classificação de futuros alunos.
* Auxiliar na disseminação da importância do uso da tecnologia computacional para a análise de dados e propor um maior contato com a mesma para testes futuros.
* Identificar os fatores que influenciam na evasão.
* Identificar qual algoritmo fornece dados mais precisos com base na extração e análise dos dados.
* Do ponto de vista teórico, evidenciar a importância da mineração de dados como ferramenta para traçar perfis e reduzir a taxa de desistência.

**3. JUSTIFICATIVA**

Este projeto carrega em si cunho educativo, desde então, é abordado um problema frequente o qual vem crescendo exponencialmente, tornando-o assim algo a se preocupar, já que não afeta apenas as redes de ensino superior, dessa forma, a partir do uso da técnica de mineração de dados, será possível analisar e avaliar se é possível classificar e traçar o perfil de alunos potencialmente evasores. Nesse sentido, pode-se vislumbrar a importância da análise proposta, que identifique uma plataforma mais precisa e eficiente para a mineração de dados, a fim de analisar e classificar os alunos de acordo com sua situação social e econômica, e então propor medidas de evitar ou reduzir o índice de evasão à IE referida.

**4. FINALIDADE**

Este estudo terá como a principal finalidade auxiliar na redução da alta taxa de evasão nas Instituições de Ensino, dando início no estudo de caso na UESC e assim, se possível, expandir para um estudo mais amplo e complexo que alcance os mais altos níveis educacionais, auxiliando na identificação.

**5. REFERENCIAL TEÓRICO**

A metodologia a qual será utilizada neste projeto é guiada pelo processo de Descoberta de conhecimento em bases de dados, do inglês Knowledge Discovery in Databases (KDD). De acordo com Fayyad (1996), KDD é um processo não trivial que objetiva identificar padrões válidos, novos (antes desconhecidos), potencialmente úteis e, essencialmente, compreensíveis em bases de dados. No processo de KDD os dados passam por várias etapas, a saber: seleção, pré-processamento, transformação e, mineração e geração de padrões, para depois estarem disponíveis como informação que pode ser visualizada e interpretada. A mineração de dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis às pessoas responsáveis pela análise dos dados (Hand et al.; 2001). Devido à ampla área de aplicação da Mineração de Dados, seu estudo é frequentemente dividido em tarefas definidas com base no resultado esperado após o processo de mineração. Para cada tarefa há diversas técnicas e métodos a serem aplicados. Em termos de aprendizagem de máquina, estas técnicas e métodos tradicionalmente são divididos em:

* Aprendizado supervisionado (preditivo): os dados para treinamento dos modelos possuem rótulos, indicando a que classe pertence as instâncias observadas. Novas classificações são feitas com base nas classes aprendidas na base de treinamento.
* Aprendizado não supervisionado (descritivo): não há rótulos nas instâncias da base de treinamento, portanto o objetivo é identificar classes, relações ou agrupamentos nos dados.

Neste projeto, o objetivo é prever a situação do aluno, portanto um exemplo de aprendizagem supervisionada. Mais especificamente, este problema é definido como classificação, que será detalhado a seguir. Classificação consiste na tentativa de aprender a generalizar um conceito com o objetivo de identificar a classe de novos dados. A base de dados para a tarefa de classificação deve conter um ou mais atributos preditivos e um atributo classe, este último é o objetivo da classificação e deve ser do tipo discreto. Neste tipo de problema utilizam-se técnicas de aprendizado supervisionado, pois a classe a ser identificada é conhecida. De acordo com Witten et al. (2011), a tarefa de classificação pode ser dividida em duas etapas. A primeira etapa é a construção do modelo com base no conjunto de dados de treinamento, no qual se assume que cada tupla pertence a uma classe definida pelo rótulo do atributo classe. O modelo resultante desta etapa é representado através de regras de classificação, árvores de decisão, fórmulas matemáticas, etc., dependendo da técnica usada para construir o modelo. Na segunda etapa, um conjunto de dados de teste é submetido ao classificador anteriormente modelado com o intuito de avaliar sua qualidade. O rótulo indicado pelo classificador é comparado com o rótulo conhecido para cada instância do conjunto de teste. O modelo é avaliado segundo métricas específicas, e caso alcance um determinado limiar, pode ser usado para classificar dados novos cujos rótulos da classe são desconhecidos. A seguir são apresentados os três algoritmos que foram escolhidos para auxiliar no estudo de caso o qual será aplicado na UESC.

**6. NAIVE BAYES**

Naive Bayes é um classificador que ganhou popularidade nos anos 90 sendo utilizado como filtro de spam. A idéia do algoritmo é utilizar o teorema de Bayes, apresentada na Equação 1, para determinar a qual classe pertence uma observação, tupla ou registro. ​

*P(A|B) = P(B|A)P(A) P(B) (1)*

como *P(x1, x2, ..., xm|ck ) = P(x1|ck ) × P(x2|ck ) ×...× P(xn|ck)*, pode-se generalizar que *P(ck |x1, x2, ..., xn|) ∼ = P(x1|ck ) × P(x2|ck ) × ... × P(xn|ck ) × P(ck)*, na qual xi é um preditor (campo da base de dados) e ck representa a classe sendo avaliada. No ­final das contas a saída do classifi­car é dada pela Equação 2, também conhecida como *Maximum a Posteriori*, a qual signifi­ca que uma observação ou tupla irá pertencer à classe que possuir a maior probabilidade.

*y = argmaxk P(ck) × YP(xi |ck ) (2)*

Segundo *Li and Li (2015)*, o Naive Bayes apresenta as seguintes vantagens: (i) possui uma base matemática sólida; (ii) é rápido; e (iii) apesar de simples, apresenta de modo geral um bom desempenho em tarefas de classi­ficação. Por esse motivo, este algoritmo é utilizado muitas vezes como base de comparação com outros algoritmos de classifi­cação.

**7. SVM**

Máquinas de Vetor de Suporte, do inglês SupportVectorMachine, é um conjunto de métodos de aprendizado supervisionado utilizados para tanto para classificação quanto para regressão. Regressão é utilizada para prever valores quantitativos (James et al.; 2013), por esse motivo, está fora do escopo deste trabalho.

A idéia no SVM é construir um hiperplano *h(x)=w.x+b*, no qual w.x é o produto escalar entre os vetores *​w e x e w ∈ X* é o vetor normal ao hiperplano. Em duas dimensões w representa o coeficiente angular da reta. Ainda na figura, a reta principal é​ *w . x + b = 0* e as linhas, ou margens, são ​*w . x + b = –1 e w . x + b = 1*, para as linhas inferiores e superiores, respectivamente, sendo que b é computado a partir da média de todos os vetores de suporte possíveis. Detalhes sobre como obter b podem ser vistos no trabalho de *Lorena and de Carvalho (2007)*. Nesse contexto, um classificador pode ser construído usando a Equação 3.

*g(x) = sgn(h(x)) = {+1, se w . x + b > 0 –1, se w . x + b < 0 (3)*

Segundo Campbell (2000), usando manipulações algébricas, a margem m de separação entre objetos pode ser obtida pela minimização de ||w||. Como não é permitido que haja dados de treinamento entre as margens, essa SVM é dita com margens rígidas. Caso o hiperplano não seja suficiente para separar as classes, então pode-se usar técnicas de aumento de dimensionalidade, assim o que não pode ser separado em 2 dimensões por uma reta, pode dependendo dos dados ser separável em 3 dimensões, e assim por diante.

Segundo Lorena and de Carvalho (2007), as SVMs são robustas diante de dados de grandes dimensões sobre os quais outras técnicas de aprendizado obtêm classificadores super ou sub ajustados. Outra vantagem é que existe somente uma configuração ótima para a SVM em seu treinamento. Essa característica é interessante frente a técnicas como as Redes Neurais Artificiais (RNAs) (Braga et al.; 2000) que são multimodais (apresentam muitos ótimos locais, dificultando a busca) em seu treinamento. Além disso, o uso de funções Kernel na não-linearização das SVMs torna o algoritmo eficiente, pois permite a construção de simples hiperplanos em um espaço de alta dimensão de forma tratável do ponto de vista computacional (Burges; 1998). Por outro lado, a principal limitação das SVMs está na sensibilidade a escolha de seus parâmetros, pois pode levar a modelos imprecisos.

**8. J48**

Como já mencionado, trata-se de uma implementação do popular algoritmo C4.5 para geração de árvores de decisão, criado por Quinlan (1993) como uma extensão do seu algoritmo anterior, o ID3. Neste algoritmo o conhecimento é representado em forma de árvore de decisão, que é uma estrutura que consiste em nós rotulados com nomes de preditores ou atributos, arcos rotulados com os possíveis valores para os atributos preditivos e folhas rotuladas com as diferentes categorias de classe. Novas instâncias são classificadas percorrendo um caminho da árvore, o que corresponde a execução de uma sequência de testes. O papel do algoritmo é escolher as regras mais importantes para compor a árvore e descartar regras que são menos adequadas. Seu funcionamento é dado por: Dada uma base de treinamento D, o algoritmo gera uma árvore inicial usando a abordagem de dividir e conquistar, utilizando as seguintes regras: (i) Se todas as instâncias em D pertencem à mesma classe ou se a base D é muito pequena, a árvore pode ser representada como uma folha rotulada com a classe mais frequente em D; (ii) Caso contrário, é escolhido um atributo com dois ou mais valores (ou faixas de valores) para ser usado como teste. Esse atributo será o nó raiz da árvore, com uma ramificação para cada possível saída do nó. A base de dados D é então dividida em subconjuntos, um para cada possível saída do nó raiz. O mesmo procedimento é aplicado recursivamente em cada um dos subconjuntos. Normalmente o algoritmo usa a entropia (H(D)) e o ganho de informação para decidir como construir a árvore. A entropia determina a impureza de um determinado conjunto de dados, sendo computada pela Equação 4, na qual PI é a proporção do atributo ser da classe i. Já o ganho de informação, dado o atributo A, que tem um domínio v é dado pela Equação 5, na qual Dj é o sub conjunto do domínio de Ae|Dj| é a quantidade de registro sem Dj .

*H(D) = –× log2pi (4)*

*GI(D, A) = H(D) – × H(Dj ) (5)*

Assim, o atributo que possuir maior ganho de informação será utilizado em um nó da árvore. O processo continua recursivamente até que a árvore esteja construída. A principal vantagem das árvores de decisão é ser compreensível à leitura humana, ou seja, é fácil de ser entendida pelo usuário. Além disso, pode lidar com uma variedade grande de dados (categórica, numérica e textual), não sendo sensível à dados faltantes (Bhargava et al.; 2013). Por outro lado, podem apresentar as seguintes desvantagens: (i) podem gerar árvores complexas, sendo sua criação computacionalmente intensa; e, (ii) dados desbalanceados podem levar a árvores grandes que são de difícil interpretação.

**9. METODOLOGIA**

Além de pesquisa bibliográfica, será utilizada a observação direta aos dados contidos no sistema acadêmico, analisando de forma estatística, qualitativa e quantitativa fatores que propiciam a evasão. A técnica de amostragem utilizada será a probabilística, pois será necessária uma análise ampla do perfil de cada aluno, tanto de ingressantes como de concludentes, a fim de traçar um parâmetro para comparação e enfim classificar os potenciais evasores. O levantamento de dados e a coleta de dados serão realizados por meio do acesso ao banco de dados da Universidade, o qual possui informações pessoais, sociais e econômicas do aluno. Os dados secundários serão obtidos por meio de levantamento bibliográfico, com consulta a materiais técnicos da área de computação, sendo eles algoritmos e banco de dados. A seleção se dará por meio do uso dos algoritmos Naive Bayes, Vector Support Machine e o J48, de modo que, o algoritmo o qual apresentar resultados mais precisos, seja tomado como base para a comparação e classificação dos alunos. A escolha de uma amostra probabilística, com tratamento de dados de cunho tanto qualitativo como quantitativo adequa-se ao tipo de pesquisa proposto, uma vez que a pesquisa não só analisa os fatores da evasão, mas também analisa a variação e crescimento da taxa de evasão, sendo necessária uma análise quantitativa desses dados.

**10. CRONOGRAMA**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FASES DA PESQUISA** | **MESES** | | | | | | | | | | | |
| **Revisão Bibliográfica** | **x** | **x** |  |  | **x** |  |  | **x** | **x** | **x** |  |  |
| **Coleta de dados** |  | **x** | **x** | **x** |  |  |  |  |  |  |  |  |
| **Aplicação de algoritmos** |  |  | **x** | **x** | **x** | **x** |  |  |  |  |  |  |
| **Comparação dos Resultados e Análise de Precisão** |  |  |  |  | **x** | **x** | **x** |  | **x** |  |  |  |
| **Análise e Discussão dos Resultados** |  |  |  |  |  | **x** | **x** | **x** |  |  |  |  |
| **Classificação dos Dados** |  |  |  |  |  | **x** | **x** | **x** | **x** | **x** | **x** |  |
| **Elaboração do Relatório** |  |  |  |  |  |  |  |  |  |  |  | **x** |

**11. ORÇAMENTO**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Discriminação** | **Quantidade** | **Valor Unitário** | **Valor Total** | **Fonte de Recurso** |
| **Computador** | **1** | **R$ 5.500,00** | **R$ 5.500,00** | **UESC** |
| **Diária** | **7** | **R$ 100,00** | **R$ 700,00** | **Próprio** |
| **Material de consumo** | **-** | **R$ 1.200,00** | **R$ 1.200,00** | **FAPESB** |
| **Total próprio** | **-** | **-** | **R$ 700,00** | **-** |
| **Total financiado** | **-** | **-** | **R$ 6.700,00** | **-** |
| **Total projeto** | **-** | **-** | **R$ 7.400,00** | **-** |

**12. ANÁLISE DE IMPACTOS**

Como parte da construção do conhecimento, um projeto científico promove impacto, seja ele direto ou indireto, dessa maneira, o estudo o qual será proposto neste projeto promoverá os seguintes impactos:

* Tecnológico​: o uso de algoritmos estará intrinsecamente relacionado com uma melhor organização dos bancos de dados das Instituições de Ensino, o que promoverá um melhor aproveitamento da tecnologia existente e a busca pela criação de novas plataformas mais intuitivas e de melhor compreensão.
* Humano: ​ com o avanço das tecnologias de classificação, tanto no âmbito educacional, como em qualquer área a qual esteja sujeita à utilização dos algoritmos que serão tomados como base para o estudo, se dará a necessidade de especialização humana para a manipulação computacional, além disso, acarretará na diminuição da carga horária de trabalho humano, visto que os algoritmos estariam responsáveis pela classificação e o humano apenas à manipulação dos códigos e manutenção.
* Científico: como já mencionado anteriormente, com o frequente uso da computação, terá como necessidade principal a criação de plataformas mais intuitivas e mais fáceis de compreender, de modo a facilitar a manipulação desses algoritmos, o que impactará na constante evolução dos mecanismos de estruturação e análise de banco de dados.

**13. BIBLIOGRAFIA CONSULTADA**

**BHARGAVA, N., SHARMA, G., BHARGAVA, R. and MATHURIA, M.​** (2013). Decision tree analysis on j48 algorithm for data mining, International Journal of Advanced Research in Computer Science and Software Engineering 3(6): 1114–1119.

**BRAGA, A., DE CARVALHO, A. C. P. L. F. and LUDERMIR, T. B.​** (2000). Redes Neurais Articiais: Teoria e Aplicações, LTC, Rio de Janeiro.

**BURGES, C. J. C.​** (1998). A tutorial on support vector machines for pattern recognition, Knowledge Discovery and Data Mining 2(2): 1–43.

**CAMPBELL, C.​** (2000). An introduction to kernel methods, in R. J. Howlett and L. C. Jain (eds), Radial Basis Function Networks: Design and Applications, Springer Verlag, Berlin.

**FACELI, K., LORENA, A. C., GAMA, J. and DE CARVALHO, A. C. P. L. F.** ​(2011). Inteligência Articial: Uma Abordagem de Aprendizagem de Máquina, LTC/Grupo Gen.

**FAYYAD, U.** ​(1996). From data mining to knowledge discovery in databases, AI magazine 17(3): 37–54.

**HAND, D., MANNILA, H. and SMYTH, P.​** (2001). Principles of Data Mining, The MIT Press. LI, L. and LI, C.​ (2015). Research and improvement of a spam lter based on naive bayes, 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol. 2, pp. 361–364.

**LORENA, A. C. and DE CARVALHO, A. C. P. L. F.​** (2007). Uma introdução às support vector machines, RITA XIV(2).

**OBSIVAC, T., POPENLINSKY, L., BAYER, J., GERYK, J. and BYDZOVSKA, H.** (2012). Predicting drop-out from social behaviour of students, Proceedings of the 5th International Conference on Educational Data Mining.

**WITTEN, I. H., FRANK, E. and HALL, M. A.​** (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3 edn, Morgan Kaufmann Publishers.