

Aplicação de Mineração de Dados para Avaliação da Importância das Variáveis dos Grupos de Estudo do PMI-PE para Sucesso na Certificação

Data Mining Application for Assessing the Importance of Variables in PMI-PE Study Groups for Success in Certification

Elifrancis Rodrigues Soares¹  orcid.org/0000-0002-2802-4050

Rodrigo Lins Rodrigues²  orcid.org/0000-0002-3598-5204

¹ Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,

² Pós-graduação em Engenharia de Sistemas, Escola Politécnica de Pernambuco, Pernambuco, Brasil,

E-mail do autor principal: Elifrancis Soares elifrancis.soares@gmail.com

RESUMO

Este artigo apresenta os resultados da aplicação de uma abordagem de mineração de dados educacionais para modelar "sucesso na certificação" no grupo de estudos do Capítulo Recife, Pernambuco Brazil Chapter® (PMI-PE). Atualmente o grupo de estudos prepara pessoas para obtenção da certificação PMP® e demais certificações do PMI. Durante a realização das atividades do grupo, existem vários instrumentos de acompanhamentos que são utilizados para verificar a evolução dos participantes. Com base nos dados gerados por estes instrumentos de coleta, buscamos neste trabalho identificar os principais fatores que implicam na obtenção da certificação. Os dados utilizados nesta pesquisa foram coletados nos anos de 2016 e 2017 referente aos grupos de estudos para certificação PMP®, CAPM®, PMI-ACP® e PMI-RMP®. A certificação do Project Management Professional do PMI aprimorou o profissionalismo da prática de gerenciamento de projetos, fornecendo aos gerentes de projeto uma abordagem estruturada para executar suas responsabilidades de projeto.

PALAVRAS-CHAVE: PMP; Certificação; Modelo.

ABSTRACT

This paper presents the results of applying an educational data mining approach to modeling (success in certification) in the Recife Chapter, Pernambuco Brazil Chapter (PMI-PE) study group. Currently, the study group prepares people for PMP® certification and other PMI certifications. During the activities group's, there are several accompaniment instruments that are used to verify the evolution of the participants. Based on the data generated by these collection instruments, we seek in this work to identify the main factors that imply obtaining the certification. The data used in this research were collected in the years 2016 and 2017 for the study groups for PMP®, CAPM®, PMI-ACP® and PMI-RMP® certification. PMI Project Management Professional certification has enhanced the professionalism of project management practice by giving project managers a structured approach to execute their project responsibilities.

KEY-WORDS: PMP; Certification; Model.

1 INTRODUÇÃO

A estabilidade financeira de todas as empresas depende do sucesso da gestão de recursos dedicados à conclusão de projetos dentro de orçamento predeterminado. Embora o número total de gerentes de projeto profissionais (PMP®) atualmente registrados no *Project Management Institute* (PMI®) é de 876.500 [1]. Por exemplo, o sucesso de um projeto depende da capacidade de um gerenciador de software para garantir a entrega oportuna, aderir às restrições orçamentárias e gerenciar especificações de escopo e qualidade. Um gerente de projeto pode ter habilidades necessárias para orientar uma equipe através de vários estágios e ciclos de vida do projeto, mas a motivação da equipe - algo bem menos tangível - é um elemento essencial de um projeto de sucesso [2].

O *Project Management Professional* (PMP®) é o padrão de certificação de gerenciamento de projetos universalmente aceito, embora as associações profissionais na Inglaterra e na Austrália também tenham programas de certificação. O PMP® é oferecido pelo *Project Management Institute* (PMI) de Newtown, Pensilvânia, que começou a estabelecer padrões em 1978, e tem oferecido certificação desde 1984. Os últimos anos têm visto um aumento acentuado daqueles que buscam a certificação PMP® (18.000 no ano 2000), a maioria nas áreas de construção, sistemas de informação e tecnologia [3].

2 CONTEXTO

O PMI (*Project Management Institute*), fundado em 1969 nos EUA, é considerado a maior associação mundial sem fins lucrativos, voltada ao Gerenciamento de Projetos, com mais de 500.000 membros, profissionais certificados e voluntários em praticamente todos os países do mundo a aumentar o sucesso das suas empresas, evoluir em suas carreiras e tornar a profissão mais madura [1]. No Brasil, o PMI está presente em 15 estados. A representação do PMI em Pernambuco - PMI-PE foi a primeira a ser fundada na região Nordeste, em 2003. O PMI em todo Brasil vem promovendo discussões sobre Gerenciamento de projetos e possibilitando um pensamento inovador e a troca de ideias sobre todos os aspectos do gerenciamento de projetos, suas práticas, negócios e resultados. Para isso, vem promovendo alianças,

palestras e encontros de profissionais ligados ao tema, visando capacitar executivos de todos os setores para que, reconhecendo a importância do gerenciamento de projeto e utilizando em seu cotidiano profissional, possam atingir melhores resultados e sucesso profissional.

O PMI oferece um programa completo de certificação para profissionais de projeto de todos os níveis de escolaridade e de qualificação. Atualmente constituído por seis credenciais, o programa demonstra o seu compromisso com a profissão e os seus conhecimentos, experiência e competência. A participação dos envolvidos se dá através de filiação.

O PMP® foi iniciado em 1984 "para reconhecer num indivíduo a compreensão do conhecimento e habilidades para liderar e direcionar equipes de projeto e entregar resultados dentro dos limites restrições de cronograma, orçamento e recursos financeiros" [4].

A filiação ao PMI (*Project Management Institute*) e ao PMI-PE (*PMI Recife, Pernambuco Brazil Chapter*®) é facultada a todos os interessados em Gerenciamento de Projetos. Os filiados (membros) ao Capítulo PMI-PE constituem um grupo de profissionais interessados no seu próprio desenvolvimento profissional e na evolução da profissão de Gerente de Projetos.

2.1 Grupo de estudos em Pernambuco

Em 2004, os membros do PMI-PE, a partir de suas experiências em cursos preparatórios para certificação PMP®, resolveram criar um grupo com objetivo de compartilhar conhecimento e experiências para prestar o exame de Certificação PMP®. Após essa investida de sucesso, os membros do Chapter pensaram em promover agregação de valor aos filiados e ainda ajudar na difusão dos fundamentos do PMBOK.

O PMI-PE desenhou um formato onde os próprios filiados pudessem discutir os temas e, em conjunto, contribuir para a solução de questões dos simulados preparatórios e compartilhar conhecimento das áreas do PMBOK®, sempre orientados por um facilitador mais experiente, que já houvesse obtido a certificação para a qual o grupo estava se preparando. Surgiu os Grupos de Estudos para certificação do Chapter Recife, com reuniões pré-determinadas, um facilitador para prover a construção do conhecimento e troca de

experiências acerca do exame, local de prova, tópicos específicos. Em média, foram executados, entre os anos 2004 e 2017, pelo menos 2 grupos de estudos por ano. Cada grupo conta com uma média entre 15 e 30 participantes.

Cada Grupo de estudo é formado por um Gerente de Projeto que é responsável por reunir os participantes, prover espaço físico para os encontros, selecionar facilitadores e fazer a interação com todos os envolvidos. Os facilitadores são filiados ao chapter, possuem a certificação para o mesmo foco do grupo (PMP¹, CAPM², PMI-ACP³ e PMI-RMP⁴) e tem for função guiar o grupo para melhor aproveitamento do tempo do encontro e que não haja dispersão do assunto. A seguir, um exemplo do cronograma dos encontros dos grupos de estudos.

Tabela 1 – Cronograma dos encontros dos grupos de estudos

N	Encontros
1	Reunião de Abertura (Kickoff Meeting)
2	Simulado de Sondagem Inicial
3	Caps 1,2 e 3 – Project Management Framework
4	Cap. 5 – Escopo
5	Mini simulado 1 + discussão
6	Cap. 6 – Tempo
7	Informação do Processo de Elegibilidade
8	Cap. 7 – Custos
9	Mini simulado 2 + discussão
10	Cap. 8 – Qualidade
11	Cap. 10 – Comunicação/Cap. 13 - Partes Interessadas
12	Cap. 9 – RH
13	Mini simulado 3 + discussão
14	Cap. 11 – Riscos
15	Cap. 12 – Aquisição
16	Mini simulado 4 + discussão
17	Cap. 4 – Integração + Ética e Responsabilidade Profissional
18	Simulado PMP com 200 Questões [online]
19	Discussão do Simulado
20	Reunião de Encerramento

Os encontros dos grupos têm, sempre, duração de 03 horas, distribuídas da seguinte forma:

- 30/40 Minutos para a apresentação do tema;
- 20 minutos para *coffee break*;
- 2 horas para *game*.

Este game é uma forma de revisão do conteúdo em forma de uma “competição” entre dois grupos de participantes, em que questões de simulados são apresentadas aos grupos e estes, de forma alternada têm preferência na resposta da questão. Se o grupo que tiver a preferência não acertar, o outro grupo pode tentar acertar a resposta. Vence o grupo que conseguir responder a maior quantidade de questões certas. O *coffee Break* também é utilizado como momento de aprendizado, um momento que as relações podem ser construídas e as trocas de experiência são fomentadas.

Durante o grupo de estudos, são realizados 06 (seis) encontros para realização de simulados com os temas estudados até então.

O índice de aprovação em certificação de Gerenciamento de Projeto em Pernambuco tem crescido de 2004 para hoje, a partir da implementação de grupos de estudos destinados aos candidatos que vão submeter à certificação. Com metodologias ativas dentro do processo de ensino de resoluções de questões e discussão dos principais pontos que são trabalhos em certificações, o grupo de estudo surgiu para possibilitar ao candidato a preparação e resoluções de questões mais complexas, bem como trabalhar o tempo destinado na preparação do estudo. De acordo com os registros do PMI-PE, desde o ano de 2014 até o momento tivemos 176 participantes e desse montante, apenas, 40 conseguiram passar no exame e obter a certificação, isso representa um percentual de ~23%.

Diante das informações supracitadas, há um potencial de 77% que o PMI-PE poderá investir para gerar novos certificados e como consequência ganhos de outros benefícios para o Capítulo, como por exemplo, aumento no número de filiados e voluntários.

Como benefício para os filiados do PMI-PE é oferecido o grupo de estudos para certificação, no qual existem uma grande procura e normalmente a quantidade que se inicia as turmas são em torno de 14 pessoas, mas ao longo existem abandonos e

1PMP® - Profissional em Gerenciamento de Projetos.
2 CAPM® - Profissional Técnico Certificado em Gerenciamento de Projetos.

3PMI-ACP® - Profissional Certificado em Métodos Ágeis do PMI.
4 PMI-RMP® - Profissional em Gerenciamento de Riscos do PMI.

não se tem um instrumento que possa aferir a qualidade da turma e direcionamento de ações que possam elevar o número de êxitos de participantes na prova de certificação.

3 TRABALHOS RELACIONADOS

Aproveitando o trabalho de Nonaka e Takeuchi [6], os autores enfatizam a necessidade de avaliar a gestão de aprendizagem, tal como ocorre no indivíduo, grupo e empresa (organização) níveis. Eles afirmam que, a fim de mover além da limitada aprendizagem de programas treinamento que apenas fornecem indivíduos com novos conhecimentos, há uma necessidade de “Transformar as lições aprendidas, em organizações e corporações desenvolvimento” [7]. Da mesma forma, Turner, Keegan e Crawford [8] aborda a importância da aprendizagem no sucesso do desenvolvimento profissional do gerente de projeto. Em termos de metodologias de treinamento no indivíduo, grupo e níveis de organização, várias sugestões tornaram-se relevantes para o conhecimento em gerenciamento de projetos.

Machine Learning (ML) tem sido amplamente adotado por diferentes setores, incluindo saúde, finanças, bancos, telecomunicação, educação, segurança, etc. No campo Educação, que é o foco deste trabalho, ML é aplicável para identificando padrões que categorizam certos comportamentos, realizações, avaliações e desempenho dos alunos; um passo considerado importante na tomada de decisões adequadas [5]. Os benefícios do ML no setor de educação podem ser classificados em seis categorias: (1) apoiar a decisão de carreira do aluno, (2) identificar padrões ocultos semânticos / não semânticos, (3) predição do desempenho dos estudantes, (4) exploração automática de grandes quantidades de dados, (5) organizar e construir perfis de alunos, e (6) apoiar os gerentes a entender melhor a natureza de seu o negócio [9].

A abordagem de mineração de dados educacionais para modelar (perda de status acadêmico) na Universidade Nacional de Colômbia. Dois modelos de mineração de dados foram definidos para analisar dados acadêmicos e não acadêmicos; os modelos usam duas classificações Bayes ingênua e o classificador de árvores de decisão, a fim de para adquirir uma melhor compreensão do atrito durante as primeiras matrículas e avaliar a qualidade dos dados para a tarefa de classificação, que pode ser entendida

como a previsão da perda do status acadêmico devido ao baixo desempenho acadêmico [10].

4 METODOLOGIA

Os dados utilizados nesta pesquisa provêm das atividades de análise de dados gerados através do controle e acompanhamento da evolução do participante do grupo de estudo. Todos os grupos existentes no capítulo para a preparação na obtenção da certificação PMP® e demais certificações do PMI foram acompanhados através de instrumentos e registro de sua participação e evolução, desde ficha de inscrição, como controle de frequência, monitoramento dos resultados dos simulados e avaliação periódica dos encontros do grupo de estudo.

Os métodos e técnicas quantitativas que envolveram a coleta de dados, análise e interpretação dos resultados do estudo foram: coleta rigorosa de informações para construção do banco de dados dos participantes de grupo de estudo, análise do desempenho nos simulados e obtenção da certificação, considerando o ano e tempo de preparação para a certificação.

4.1 Objetivo geral

O objetivo desta pesquisa foi a aplicação de algoritmos de mineração de dados para a identificação das variáveis mais significativas, relacionadas as ações dos grupos de estudo do PMI-PE.

4.2 Objetivos específicos

- desenvolver indicadores que possibilitem o controle e o acompanhamento de participantes;
- elencar os principais pontos de melhoria para auxiliar no crescimento do grupo de estudo;
- verificar quais variáveis tiveram maior impacto no sucesso das certificações.

O método utilizado nesta pesquisa seguiu o processo CRISP-DM (*Cross Industry Standard Process for Data Mining*). O CRISP-DM é um padrão industrial originalmente desenvolvido pela NCR Systems Engineering Copenhagen, Daimler Chrysler AG, SPSS Inc. e OHRA [11].

Para os propósitos deste trabalho, instanciamos o processo CRISP-DM em um ambiente

educacional, adaptando suas seis fases da seguinte forma: compreensão do domínio educacional a modelar, compreensão dos dados educacionais envolvidos, preparação de dados para a aplicação de técnicas EDM, modelagem, avaliação modelo e implementação da solução educacional. A Figura 3 mostra essas fases e suas inter-relações.

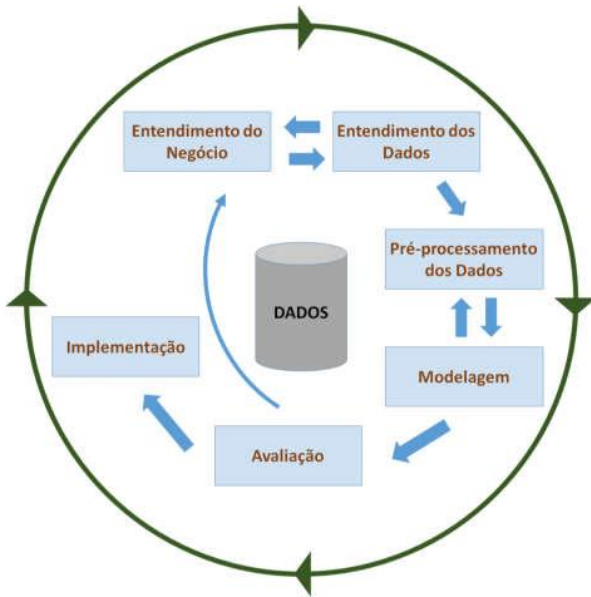


Figura 1: Ciclo da Metodologia CRISP-DM

Árvore de Decisão é uma das técnicas mais populares de mineração de dados. Mais comumente usada para resolver a tarefa de classificação de dados, a árvore de decisão consiste em uma coleção de nós internos e nós folhas, organizados em um modelo hierárquico (da mesma forma que se organizam as estruturas de dados do tipo de árvore). No contexto da resolução da tarefa de classificação, uma árvore de decisão representa o modelo capaz de guiar a tomada de decisão sobre a determinação da classe à qual pertence.

Há uma série de pacotes que implementam árvores de decisão em R. Adicionalmente, a árvore resultante como modelo classificador também pode ser interpretada de forma gráfica. Portanto, há também pacotes específicos para a visualização do resultado. Para esta pesquisa, usamos a função árvore de decisão `rpart()`, disponível no pacote `rpart` [12]. O significado do nome da função `rpart()` vem de *Recursive Partitioning and Regression Trees*.

Os parâmetros solicitados pela função são descritos abaixo:

- **Fórmula** define qual é o atributo de classe e quais são os atributos descritivos do conjunto de dados. Por exemplo: $y \sim x_{11} + x_{12} + x_{13}$;
- **Dados** é um conjunto de dados armazenado como um `data.frame`;
- **Método** é o parâmetro que define se árvore será usada como classificação ("class") ou regressão ("anova");
- **Controle** é a variável na qual são armazenados os parâmetros que controlam o crescimento da árvore;
- **Parâmetros** é a variável na qual é definido o critério de seleção de atributos.

A função retorna um objeto de árvore de decisão.

```
modelo_ad <- rpart(Certificado ~ Renovou + Avaliacao +
  Simulados + Quant_Simulado + Presenca + Dias_Certificado_GE,
  data = treinamento, method = "class", control =
  rpart.control(minsplit = 1), parms = list(split = "Information"))
```

4.3 Entendimento dos Dados

Os dados coletados para esta pesquisa compreendem os anos de 2016 a 2017, e foram fornecidos pelos grupos de estudo para certificação PMP®, CAPM®, PMI-ACP® e PMI-RMP®. Foi realizando um levantamento dos diversos arquivos dos grupos de estudos, os quais podem ser classificados

- Avaliações dos Encontros;
- Cronograma do Grupo de Estudos;
- Performance nos Simulados;
- Presença.

Foram coletados dados de avaliações realizadas após cada encontro do grupo de estudo, nos quais o participante pôde responder um formulário online com o de identificar possíveis melhorias ao longo de todo o processo.

Outro conjunto de dados variável coletado foram os resultados dos simulados. Estes simulados são opcionais e aplicados em momentos específicos para que o participante possa aferir seus conhecimentos ao longo do curso.

A variável alvo para este estudo foi a obtenção de certificação pelo estudante. É uma variável binária que pode receber o valor *True* ou *False*, ou seja, se o aluno conseguiu ter êxito na certificação ou não. A técnica empregada foi o aprendizado supervisionado, pois foi possível ter uma base de dados com o rótulo referente a certificação dos estudantes

Abaixo são listadas as variáveis utilizadas no estudo:

- **Renovou** - Se o participante renovou a filiação depois de um ano;
- **Avaliacao** - Média do resultado das avaliações;
- **Simulados** - Média dos resultados dos simulados;
- **Quant_Simulado** - Quantidade de simulados que o participante realizou;
- **Presenca** - Índice de presença do participante;
- **Dias_Certificado_GE** - Diferença de dias entre a data da certificação e a data de finalização do grupo de estudo;
- **Certificado** - Se o participante teve a certificação que foi o objetivo do grupo de estudo.

Na tabela 1 temos o resumo dos dados, no qual podemos observar os valores mínimo, máximo, média, mediana e desvio padrão das variáveis selecionadas. Atenção para os valores da média e mediana da variável simulado que é baixo, como também a variável Quant_Simulado, mostra que é a ação de aplicação de simulados pouca utilizada pelos participantes do grupo de estudos.

Tabela 1: Resumo dos dados

Renovou	Avaliacao	Simulados	Quant_Simulado
Min. :0.0000	Min. :0.000	Min. :0.0000	Min. :0.000
Max. :1.0000	Max. :5.000	Max. :1.9300	Max. :6.000
Mean :0.3313	Mean :2.814	Mean :0.3234	Mean :2.805
Median :0.0000	Median :4.150	Median :0.2400	Median :3.000
SD: 0.4721013	SD: 2.198193	SD: 0.3340657	SD: 1.746691
Presenca	Dias_Certificado_GE	Certificado	
Min. :0.0600	Min. : 0.00	Min. :0.0000	
Max. :1.0000	Max. :373.00	Max. :1.0000	
Mean :0.7265	Mean :29.04	Mean :0.2722	
Median :0.8000	Median : 0.00	Median :0.0000	
SD: 0.2344553	SD: 57.70021	SD: 0.4464096	

A figura 4 mostra os resultados da análise no qual pode-se observar que a variável Dias_Certificado_GE é a que apresenta maior correlação com a variável alvo Certificado, depois temos a correlação quase moderada das variáveis Quant_Simulado e Simulados. Um ponto que chama atenção é que variável Presenca possui baixa correlação com Certificado, o que sugere que a frequência pode não influenciar na obtenção da certificação.

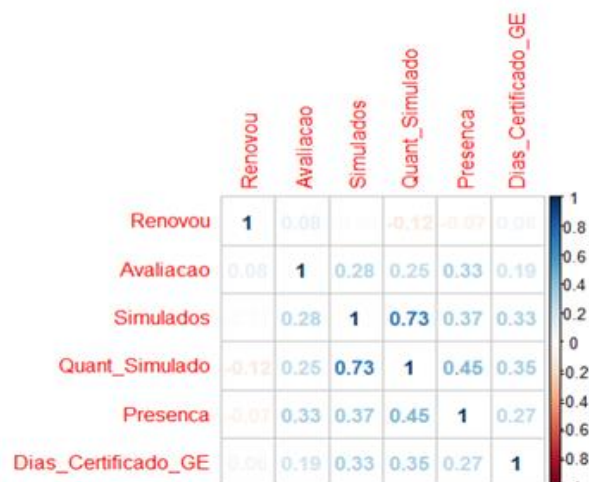


Figura 2: Correlação entre variáveis

Na figura 2, podemos observar a importância da variável Quant_Simulado, mas mesmo assim com baixa correlação. A variável Simulados é a segunda mais importante com 0.33 de correlação. O ponto que chama atenção é a variável presença que ficou correlação muito fraca.

4.4 Pré processamento dos dados

A extração e limpeza dos dados foi uma das fases que exigiu maior esforço. Foram criadas novas variáveis a partir dos dados coletados, o que enriqueceu os dados e pôde nos ajudar a melhorar o resultado final do modelo. Nesta fase foi realizada a remoção de algumas variáveis, por não se mostraram significativas para a construção do modelo. No início haviam 39 variáveis, e após restaram 7 variáveis.

As variáveis criadas foram:

- **Avaliacao** - Realizamos a média do resultado das avaliações, no qual tínhamos 11 variáveis (AC_01, AC_02, AC_03, AC_04, AC_05, AT_01, AT_02, AT_03, AT_04, AM_01 e AM_02);
- **Simulados** - Realizamos a média dos resultados dos simulados, no qual tínhamos 6 variáveis (Simulado_Sondagem, Simulado_I, Simulado_II, Simulado_III, Simulado_IV e Simulado_Final);
- **Quant_Simulado** - Criamos esta variável para saber o quantitativo de simulados realizados por cada participante;
- **Dias_Certificado_GE** - Foi calculado a diferença de dias entre a data da

certificação e a data de finalização do grupo de estudo.

Algumas outras variáveis foram removidas, pois não apresentavam significância para a construção do modelo. Segue a listagem de todas variáveis removidas:

- AC_01 - Avaliação do encontro - Pergunta: Aspectos Comportamentais - Mostrou-se solícito e proativo;
- AC_02 - Avaliação do encontro - Pergunta: Aspectos Comportamentais - Gerenciou bem o tempo;
- AC_03 - Avaliação do encontro - Pergunta: Aspectos Comportamentais - Transmitiu o conhecimento com clareza;
- AC_04 - Avaliação do encontro - Pergunta: Aspectos Comportamentais - Promoveu a interação do grupo;
- AC_05 - Avaliação do encontro - Pergunta: Aspectos Comportamentais - Promoveu discussões e debates;
- AT_01 - Avaliação do encontro - Pergunta: Aspectos Técnicos - Demonstrou conhecimento e domínio sobre o assunto;
- AT_02 - Avaliação do encontro - Pergunta: Aspectos Técnicos - Apresentou bom número de exemplos durante a explicação;
- AT_03 - Avaliação do encontro - Pergunta: Aspectos Técnicos - Mostrou dicas sobre o exame de certificação;
- AT_04 - Avaliação do encontro - Pergunta: Aspectos Técnicos - Alinhamento com o GAME;
- AM_01 - Avaliação do encontro - Pergunta: Ambiente e Material - Conforto do espaço físico (ar condicionado, data show, bancada);
- AM_02 - Avaliação do encontro - Pergunta: Ambiente e Material - Nível de ruído da sala;
- Simulado_Sondagem - Resultado do simulado de sondagem do participante antes de iniciar o grupo de estudo;
- Simulado_I - Resultado do primeiro simulado depois que iniciou o grupo de estudo;
- Simulado_II - Resultado do segundo simulado depois que iniciou o grupo de estudo;

- Simulado_III - Resultado do terceiro simulado depois que iniciou o grupo de estudo;
- Simulado_IV - Resultado do quarto simulado depois que iniciou o grupo de estudo;
- Simulado_Final - Resultado do simulado final depois que iniciou o grupo de estudo
- Finalizou - Se o participante finalizou o grupo de estudo;
- Fim_Turma - Data da finalização do grupo de estudo;
- Data_Certificacao - Data que a participante tirou a certificação.

4.5 Modelagem

Para construir os modelos, escolhemos a função de classificação. A classificação é o processo de atribuição de itens de uma coleção às classes alvo, usando um conjunto de itens pré-classificados ou dados de treinamento [13].

Utilizamos a linguagem de programação R⁵ para implementar quatro modelos de classificação. O primeiro foi baseado Árvores de Decisão, o segundo foi Naive Bayes, o terceiro usado foi Regressão Logística, e o quarto usou o algoritmo Suporte Vector Machines (SVM). Na tabela 2, temos os algoritmos utilizados no artigo.

Tabela 2: Algoritmos dos modelos

Modelo	Algoritmo
Arvore Decisao	<code>modelo_ad <- rpart(Certificado ~ Renovou + Avaliacao + Simulados + Quant_Simulado + Presenca + Dias_Certificado_GE, data = treinamento, method = "class", control = rpart.control(minsplit = 1), parms = list(split = "Information"))</code>
Naives Bayes	<code>modelo_nb = naiveBayes(Certificado ~ ., treinamento)</code>
Regressao Logistica	<code>modelo_rl <- glm(Certificado ~ Renovou + Avaliacao + Simulados + Quant_Simulado + Presenca + Dias_Certificado_GE, data=treinamento)</code>
SVM	<code>modelo_svm = svm(Certificado ~ Renovou + Avaliacao + Simulados + Quant_Simulado + Presenca + Dias_Certificado_GE, treinamento)</code>

Para treinar e testar os modelos desenvolvidos, dividimos o conjunto de dados em dois subconjuntos, como mostrado na Figura 3. O primeiro subconjunto compõe 70% do conjunto de dados e foi usado para treinamento e o segundo subconjunto representa 30% do conjunto de dados e foi reservado para testes.

5 <https://www.r-project.org/>

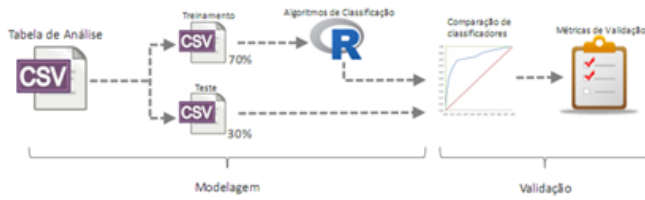


Figura 3: Processo de Treinamento e Testes

Com base nos resultados obtidos ao executar os modelos no subconjunto de teste, comparamos o desempenho de cada um dos modelos de classificação que construímos. A próxima subseção descreve os resultados obtidos.

4.6 Avaliação

O processo de avaliação visa validar os modelos de classificação testando-os com um conjunto de dados que não foi usado para treinamento. Este é um passo muito importante, no qual podemos entender as características e forças de cada modelo antes de implantá-los em um ambiente de produção [14]. Para realizar esta avaliação, utilizamos o subconjunto de treinamento que continha 119 registros.

Através da métrica de avaliação "Accuracy", o classificador Árvores de Decisão apresentou o melhor resultado (0.98) seguido de Naive Bayes e Regressão Logística (0.94), o último foi SVM (0.9).

Para escolher o melhor classificador, usamos a curva ROC [15]. Este gráfico ajuda a visualizar o desempenho de um classificador binário e também denota as variações de sensibilidade e especificidade para diferentes valores de corte [15]. A Figura 4 mostra a curva ROC para os quatro classificadores desenvolvidos neste trabalho.

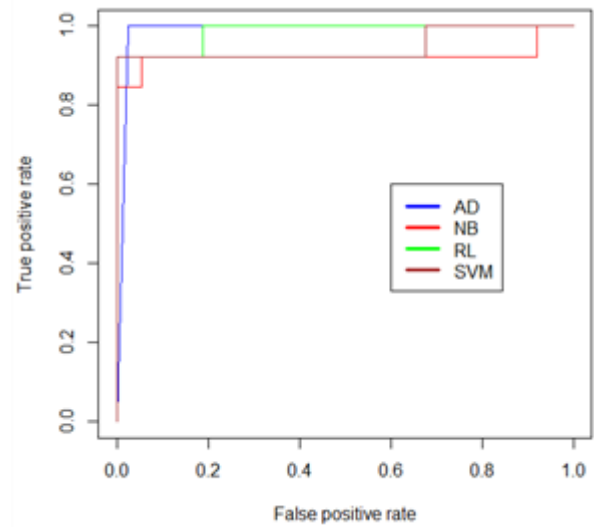


Figura 4: Comparação entre classificadores usando as curvas ROC.

5 RESULTADOS

Dado os bons resultados obtidos com o modelo de Árvore de Decisão, apresentamos uma análise detalhada do processo de modelagem usado para construir este classificador, bem como os detalhes sobre a validação e ajustes de parâmetros.

Nosso modelo de classificação foi construído a partir de um subconjunto de treinamento de 119 registros, o que corresponde a 70% do conjunto de dados completo.

A construção da árvore é realizada por meio de um algoritmo que iterativamente analisa os atributos descritivos de um conjunto de dados previamente rotulado. Podemos observar na Figura 5 que pessoas que tiveram um maior tempo de preparação conseguiram êxito na prova, mas em contrapartida a variável avaliação não se mostrou com grande significância para o sucesso na certificação.

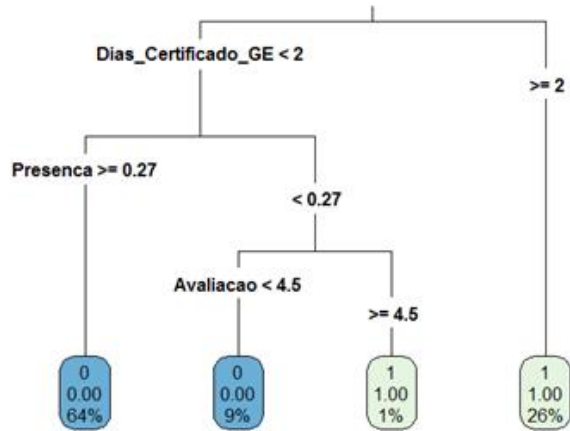


Figura 5: Árvore de decisão.

Avaliamos o desempenho do nosso modelo usando um conjunto de medidas diferentes, com base em um subconjunto de teste contendo 50 registros, o que corresponde a 30% do conjunto de dados completo. A Tabela 2 mostra os resultados desta avaliação.

Tabela 2: Matriz de Confusão e Estatísticas

```

classificacaoProb_ad  0  1
                    0 37  1
                    1  0 12

Accuracy : 0.98
95% CI : (0.8935, 0.9995)
No Information Rate : 0.74
P-value [Acc > NIR] : 5.375e-06

Kappa : 0.9467
McNemar's Test P-Value : 1

Sensitivity : 1.0000
Specificity : 0.9231
Pos Pred Value : 0.9737
Neg Pred Value : 1.0000
Prevalence : 0.7400
Detection Rate : 0.7400
Detection Prevalence : 0.7600
Balanced Accuracy : 0.9615

'Positive' Class : 0
    
```

De acordo com a matriz de confusão apresentada na Tabela 2, nosso classificador foi capaz de fazer uma predição correta em 98% dos casos, o que pode ser considerado um ótimo nível de desempenho de acordo com os limiares propostos por Faveiro, Belfiore e outros. (2009), que afirmam que um "bom modelo" faz uma predição correta em mais de 85% dos casos. Além da precisão, também usamos outras quatro

medidas para avaliar o nosso modelo: Especificidade, accuracy e área sob a curva ROC.

Para a medida de especificidade, nosso modelo atingiu o valor de 0.9231 (92,31%). Esta medida é definida como o número de instâncias negativas corretamente previstas pelo classificador, ou seja, o número de alunos que foram corretamente preditos como não em risco. Assim, podemos dizer que dos 37 alunos do subconjunto que estavam em situação de desempenho satisfatória, 13 deles foram corretamente classificados pelo modelo de árvore de decisão.

Para a medida de Accuracy, nosso modelo atingiu o valor de 0.98 (98%). Esta medida determina a porcentagem de registros corretamente identificados pelo classificador na classe positiva. Quanto maior a taxa de precisão, menor o número de erros falsos positivos cometidos pelo classificador, ou seja, quanto maior a precisão, menor será o número de participantes em risco, mas o classificador os classificou na classe de "sucesso na certificação".

Para a última medida, a Área sob a curva ROC, nosso modelo atingiu o valor de 0.98. O gráfico da curva ROC é construído com base na taxa de positivos verdadeiros e na taxa de falsos positivos - os falsos positivos são traçados no eixo x e positivos verdadeiros no eixo dos y. A Figura 6 mostra o gráfico que representa a área sob a curva ROC para o nosso modelo.

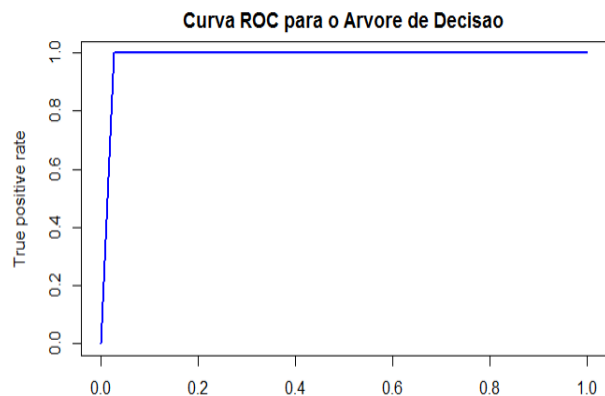


Figura 6: Curva ROC

A área sob a curva ROC é uma porção da área do quadrado da unidade (ROC), cujos valores variam de 0 a 1. Quanto maior a área abaixo da curva, melhor será o desempenho geral do classificador.

Assim, de acordo com a taxa de precisão positiva, o modelo de arvore de decisão foi o

melhor entre os quatro classificadores utilizados e avaliados nesta pesquisa.

6 DISCUSSÕES

O primeiro passo para aumentar o sucesso na certificação é a identificação de participantes com risco de mostrar desempenho ruim e ausência nos encontros. Com o uso de métodos de modelagem preditiva, é possível identificar esses participantes e seus comportamentos.

A análise preditiva permite que facilitadores e gerentes tomem medidas preventivas, pois podem identificar as ações de potenciais participantes em situação crítica. Com a criação do modelo poderá ser aplicadas análises preditiva com os dados dos novos participantes.

O grupo de estudos vem buscando continuamente melhorar, afim de aumentar o número de participantes que tenham êxito na obtenção da certificação. Atualmente existe um controle de presença, monitoramento dos resultados dos simulados, cronograma e avaliação dos encontros e facilitadores, são muitos controles paralelos e que algumas vezes acabam não sendo realizados no momento adequado, gerando um retrabalho e perda de oportunidades para uma ação imediato no direcionamento do grupo de estudos.

Foi realizado um trabalho de coleta das informações dos anos de 2016 e 2017 e organizadas de acordo com a turma e ano realizado do grupo de estudo, e proposto melhorias no processo, por exemplo: é criado um formulário para cada encontro, na geração de um acompanhamento geral da qualidade do grupo de estudos, pois é gerado uma planilha por encontro.

Um problema encontrado é que a avaliação dos encontros não obrigatória a identificação da pessoa, logo fica inviável saber o grau de satisfação do participante. O controle de presença é na sua maioria realizado no termino do grupo de estudos, gerando uma dificuldade no monitoramento continuo do percentual de presença dos participantes para que se tenha alguma ação para normalização do percentual de presença dos participantes.

Os simulados são realizados nos momentos adequados, mas é opcional o participante realizar e podendo realizar em outro sem ter o registro do percentual de acerto, está ação acaba gerando que muitos participantes deixam para realizar o simulado após o término do grupo de estudos,

gerando a falta do registro do percentual de acerto do participante nos simulados.

Não é mantido o controle da pontuação do game, logo não foi possível inserir na análise estatística deste projeto. Por fim os facilitadores não são fixos por tema do encontro e a cada encontro é na sua maioria um facilitador diferente, ficando difícil medir o grau de impacto se determinado facilitador é crucial no grupo de estudos, bem como não tem mapeado nos simulados as perguntas relacionado com o tema discutido com o facilitador.

7 CONCLUSÕES

Os grupos de estudo do PMI-PE, têm se mostrado um agregador de valor para os filiados bem como para a comunidade de gerenciamento de projetos das regiões que compreendem a área de atuação do chapter, tendo recebido reconhecimento mundial do PMI e sendo copiado para os demais capítulos do PMI. O fortalecimento dos grupos e os processos de melhoria contínua adotados pelos voluntários na busca pela excelência fazem dos grupos de estudos referência e credibilidade na preparação para as certificações PMI.

Este artigo teve como objetivo aplicar uma abordagem de mineração de dados educacionais para modelar o sucesso na certificação no grupo de estudos do Capítulo Recife, Pernambuco Brazil Chapter® (PMI-PE). Para atingir esse objetivo, utilizamos técnicas de mineração de dados para construir um modelo preditivo capaz de prever o sucesso na certificação dos participantes do grupo de estudos com base nos controles existentes. O modelo desenvolvido produziu resultados satisfatórios em experiências com dados reais de 169 participantes.

As contribuições desta pesquisa são duas: apresentamos e validamos um modelo para prever o sucesso do participante e também propomos ações de melhorias da forma de monitorar e acompanhar a evolução da execução do grupo de estudo.

Para trabalho futuro poderá ser desenvolvido a mineração dos dados e análise preditiva a respeito da filiação ao PMI-PE, pois existem um baixo nível de retenção dos filiados. Temos os dados coletados da filiação, evolução do filiado, trabalho voluntário e participação dos eventos realizados no capítulo. Este trabalho faz parte de um trabalho de organização dos dados do chapter referente às

diversas áreas (financeiro, eventos, educação, filiação, voluntariado), estruturando e criando análises.

REFERÊNCIAS

[1] Project Management Institute. Disponível em: www.pmi.org. Acessado em: nov. 2018.

[2] SCHMID, B.; ADAMS, J. Motivação no gerenciamento de projetos: a perspectiva do gerente de projetos. **Project Management Journal**, v. 39, n. 2, p. 60-71, 2008.

[3] FOTI, R. O caso da certificação. **PM Network**, v. 15, n. 9, p. 46-49, 2001.

[4] Project Management Institute. **Project Management Professional (PMP®) credential handbook**. Newtown Square, PA: Author, 2009.

[5] NONAKA, I.; TAKEUCHI, H. **The knowledge-creating company: How Japanese companies create the dynamics of innovation**. Oxford, UK: Oxford University Press, 1995.

[6] MOHD, M. A. Role of data mining in education sector. **International Journal of Computer Science and Mobile Computing**, v. 2, n. 4, p. 374-383, 2013.

[7] BERGGREN, C.; SÖDERLUND, J. Rethinking project management education: Social twists and knowledge co-production. **International Journal of Project Management**, v. 26, n. 3, p. 286-296, 2008.

[8] TURNER, J. R.; KEEGAN, A. E.; CRAWFORD, L. H. Delivering improved project management maturity through experiential learning. In: TURNER, J. R. (ed.), **People in project management**. Aldershot, UK: Gower, 2003. p.45-63.

[9] LUAN, J. Data mining and its applications in higher education. **Journal of New Directions for Institutional Research**, v. 113, p. 17-36, 2002. DOI: <https://doi.org/10.1002/ir.35>. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.35>.

[10] GUARÍN, C.; GUZMÁN, E.; GONZÁLEZ, F. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. **IEEE Revista Iberoamericana de tecnologías del Aprendizaje**, v. 10, n. 3, p. 119-125, 2015. DOI: 10.1109/RITA.2015.2452632. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7156098/>

[11] FAVEIRO, L. *et al.* **Análise de dados: modelagem multivariada para tomada de decisão**. São Paulo: Campus, 2009.

[12] THERNEAU, T.; ATKINSON, B.; RIPLEY, B. rpart: Recursive Partitioning and Regression Trees, 2014. (Version 4.1-8). Retrieved from Available at CRAN. **R-project.org/package=rpart**, 2015. Disponível em: <http://CRAN.R-project.org/package=rpart>,2018.

[13] SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados com Aplicações em R**. Rio de Janeiro: Elsevier, 2016.

[14] CHAPMAN, P. *et al.* CRISP-DM 1.0 Step-by-step data mining guide. **SPSS inc**, v. 9, p. 13, 2000.

[15] FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, v. 27, n.8, p. 861-874, 2006. DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X>.