Árvore de Decisão como Ferramenta de Inovação em Gestão da Saúde: Consumo de Álcool dos Discentes do IFBA/*Campus* Salvador

Resumo

O consumo de álcool é um problema complexo, sendo uma grave e crescente temática de saúde pública. No Brasil, esse aumento tem sido considerável, inclusive no ambiente educacional. Pesquisas apontam que, em 2006, 13% da população com menos de 15 anos tinham experimentado bebidas alcoólicas; em 2012, este número passou para 22%. Gerenciar essa tendência preocupante e ascendente, no âmbito institucional, deve ser uma das prioridades dos gestores de saúde públicos e privados, fazendo uso de informações relevantes. Resultados iniciais sobre o consumo de álcool no Instituto Federal da Bahia (IFBA) foram apresentados em eventos de pesquisa em 2018, com os dados obtidos a partir da aplicação do questionário *Alcohol Use Disorders Identification Test* (AUDIT), utilizado como referência para identificação de problemas relacionados ao uso do álcool, ao evidenciar o correspondente grau de consumo pelo indivíduo, bem como desordens associadas e possível dependência. O presente trabalho utilizou a Árvore de Decisão como forma de representação do conhecimento do comportamento do consumo ou não de álcool a partir de um extrato do banco de dados original, composto por uma amostra de 256 estudantes do IFBA/*Campus* Salvador. Com até seis perguntas, o profissional de educação ou de saúde pode prever esse comportamento sem colocar o discente numa posição de constrangimento com pergunta direta sobre consumo ou não de álcool.

Palavras-chaves: Classificação; Consumo de Álcool; Representação do Conhecimento; Gestão da Informação; Árvore de Decisão.

1. Contextualização

O interesse em atuar na prevenção do uso de drogas em ambiente acadêmico levou alguns pesquisadores liderado pelo Prof. Dr. José Lamartine a elaborar uma pesquisa diagnóstica ampla sobre as características e disseminação do seu uso no IFBA.

Apropriando-se de uma metodologia já consagrada internacionalmente, foi elaborado um questionário com 153 questões, portanto trazendo implicitamente um grau de relativa complexidade, sendo obtidos 619 registros que, a despeito da dificuldade de coleta e análise de dados, trouxe uma riqueza de informações que podem subsidiar uma ampla gama de ações gerenciais e sociais mitigadoras do consumo de droga em nossa Instituição e na sociedade como um todo, por extensão.

Visando aprofundar o entendimento da influência das diversas variáveis, o Grupo de Pesquisa de Estudo e Pesquisa de Inovação em Organizações (GEPIO) tem buscado desenvolver estudos e pesquisas que apliquem metodologias para um melhor domínio do banco de dados disponível. Assim, esse apresentou o artigo Análise de Agrupamentos como Ferramenta de Inovação em Gestão da Saúde: Consumo de Álcool dos Discentes

do IFBA/Campus Salvador[1] no CONVIBRA (2020) e, agora, vem completar o estudo sobre consumo do álcool ao utilizar técnicas de previsão de comportamento, ao gerar um classificador.

Segundo Westphal e Blaxton (1998), a visualização pode ser uma poderosa ferramenta de análise de tendência e de descoberta de padrões que podem não ser percebidas em métodos não-visuais. Com isto, a visualização permite descobrir pequenos padrões escondidos nos dados. A partir destes pontos de vista, aqueles autores caracterizam a visualização como uma oportunidade do pesquisador realizar observações dinâmicas sem preconceito, além de possibilitar análise e inspeção de dados em larga quantidade de uma só vez.

Análises complexas demandam alta capacidade cognitiva. No entanto, a restrição de desempenho da memória de curto prazo em lidar como novos itens simultaneamente e ou/ a da recuperação de informações da memória de longo prazo pode se tornar fator limitante na análise e descoberta de padrões (WESTPHAL, BLAXTON, 1998). Por isso, o pesquisador tende em adotar técnicas que facilitem a compreensão dos dados, entre as quais: agrupamento, rede auto-organizáveis, georreferenciamento e a hierarquia, neste artigo, representada pelo classificador Árvore de Decisão, a ser detalhado adiante.

Para tanto, os autores esclareceram alguns temas, entre as quais: Descoberta de Conhecimento em Base de Dados, Classificação, Árvore de Decisão e Critério de Generalização do Modelo.

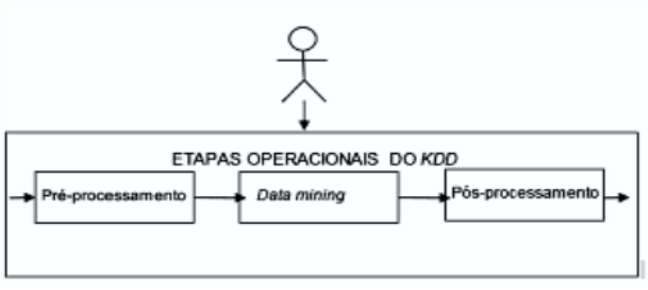
2. Descoberta de Conhecimento em Base de Dados

A Descoberta de Conhecimento em Base de Dados do inglês *Knowledge-Discovery in Database (KDD)* é definido por Goldschmidt e Passos (2005, p.3) como:

> "KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para a identificação de padrões compreensíveis, válidas, novos e potencialmente úteis a partir de grande base de dados."

Os autores ilustram as etapas do processo de KDD na Figura 1, sendo formado pelas seguintes etapas: Pré-processamento, Mineração de Dados e Pós-processamento.

Figura 1 - Etapas Operacionais do KDD segundo Goldschmidt e Passos



Fonte: Goldschmidt e Passos (2005).

[1] Disponível em: < https://convibra.org/congresso/res/uploads/pdf/artigo23685_20202835.pdf>. Acesso em: 15 set. 2022.

A etapa de Pré-Processamento consiste em identificar e coletar os dados de interesse, sendo composta por várias tarefas, entre as quais: organização dos dados num único local, eliminação dos exemplares[2] repetidos e/ou valores discrepantes (*outliers*), seleção de atributos discrepantes, normalização de atributos numéricos numa mesma escala, transformação de valores ou tipo de dados. Tais tarefas adequam os dados para serem tratados pela próxima etapa, mineração de dados.

A etapa de Mineração de Dados, conhecida como *Data Mining*, é definida como o processo de Descoberta de Padrões nos Dados, segundo Witten e Frank (2000). Silva *et al* (2016, p. 10) corrobora que esta percepção de mineração de dados pode ser definida como:

> [...] como um processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados e que sejam importantes para embasar a assimilação de informação importante, suportando a geração do conhecimento.

Berson e Smith (1997) reforçam tais declarações ao afirmar que a mineração de dados ajuda o usuário final a extrair informações úteis de negócio a partir de uma grande base de dados. Provost e Fawcett (2016) corroboram com a visão de Berson e Smith (1997) ao citarem que a mineração de dados enfoca a busca de conhecimento, padrões ou regularidades dos dados, de forma automática. Goldschmidt e Passos (2005) contextualizam a mineração de dados com uma das etapas do processo de KDD.

A mineração de dados pode ser estruturada pelas seguintes tarefas básicas: Agrupamento, Associação, Regressão e Classificação. A tarefa de Agrupamento procura formar grupos em que elementos intragrupos são similares e os elementos intergrupos são dissimilares. A tarefa Associação constrói regras de relacionamento entre os atributos, chegando a uma relação de implicação. A tarefa de Regressão procura predizer valores numéricos a partir de outros atributos. A tarefa Classificação categoriza um novo elemento a partir de outros atributos.

A etapa de Pós-Processamento se preocupa em apresentar os resultados de maneira legível, principalmente para o público não familiarizado. Assim, o resultado do processo KDD pode contribuir na resolução de alguma questão importante.

2.1. Classificação

Silva *et al* (2016, p. 11) denomina classificação como:

> "um processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence um exemplar de um domínio sob análise, com base num conjunto de dados classificados".

Westphal e Blaxton definem a classificação a tarefa de rotular os registros. Assim, a classificação é um tipo de tarefa onde um atributo não-numérico é considerado rótulo, atributo alvo, e que procura encontrar padrões de comportamento associados a outros atributos. A partir do atributo alvo, segmenta-se a base de dados. Por utilizar o atributo alvo, o aprendizado utilizado é o supervisionado, diferente do aprendizado não

[2] O mesmo de um registro de banco de dados ou uma ocorrência ou uma linha de uma tabela

supervisionado do agrupamento por não existir um atributo alvo. A classificação difere da regressão por predizer um valor numérico, enquanto aquele prediz uma categoria (PROVOST, FAWCETT, 2016). Logo, a classificação de registro pode ser implementada por classes de métodos, entre os quais: Redes Neuronais e Árvore de Decisão. Este artigo detalhará a Árvore de Decisão logo em seguida.

2.1.1. Arvore de Decisão

Segundo Berson e Smith (1997, p.351), "*Decison Tree is predictive model that, as name implies, can be view as tree*.". Já Westphal e Blaxton (1998, p. 181) conceitua como:

> "*Decision Trees are analythical tools used to discover rules and relationships by systematically breaking drown and subdividing the information contained in your data set*".

Corroborando com as citações acima, a Árvore de Decisão é um diagrama em forma de arvore invertida, utilizada na descoberta de regras e relacionamentos. Sendo assim, significa que a raiz, o nó inicial, situa-se na parte de cima do diagrama, as folhas, os nós finais, no final do diagrama, e entre a raiz e as folhas, existem os nós intermediários. Os nós são ligados pelos ramos. Os ramos armazenam as regras de divisão da base de dados, conhecida como quebra da árvore.

O nó do tipo raiz armazena todos os exemplares do conjunto de dados a ser tratado. A partir dele, realiza a quebra da árvore, subdividindo a base de dados com o objetivo de formar subconjunto com exemplares internos o mais homogêneo possível com relação ao atributo da classe-alvo[3]. Cada subconjunto de dados é derivado recursivamente até chegar a um único exemplar ou exemplares pertencentes a uma mesma classe do atributo-alvo. Com isso, a "árvore" pode predizer a classe de um novo exemplar ao aplicar as regras contidas nos ramos da árvore, configurando-se como uma "técnica de classificação".

Segundo Berson e Smith (1997), a Árvore de Decisão pode ser utilizada na:

- Exploração de dados, ao buscar preditores e valores na formação dos subconjuntos, denominada de quebra da árvore em ramos;
- Pré-processamento de dados ao formar subconjuntos de dados a ser submetido a algum outro algoritmo de predição; e
- Predição ao classificar um novo elemento com base na classe do atributo-alvo (classe-alvo).

Pelo fato de se ter facilidade de entendimento de sua estrutura e possibilidade de geração de regras de classificação, a Árvore de Decisão tem uma forte adesão de uso, segundo critérios de Automação, Clareza e ROI (*Return of Investment*) de acordo com Berson e Smith (1997). Ademais, O algoritmo de implementação da Árvore de Decisão pode estar sujeito as métricas de geração dos subconjuntos contidas no Quadro 1.

[3] Atributo que identifica a classe do exemplar

Quadro 1 - 1 Métricas de Avaliação de uma Árvore de Decisão

Critério	Resumo
Ganho de Informação	Usa a medida de entropia, da Teoria da Informação, como meio para análise do grau de impureza das partições geradas a partir da análise dos valores de um atributo descritivo
Índice Gini	Usa o critério baseado em impureza para analisar as diferenças entre as distribuições de probabilidade dos valores dos atributos de classe(rótulo).
Likelihood Ration Chi-Squared Statistics	Mede a significância estatística do critério do ganho de informação
DKM	Critério baseado em impureza, especialmente, projetado para atributos de classes binários. Gera árvores menores que o ganho de informação e o índice Gini
Raio de ganho	Normaliza o critério de ganho de informação. Deve ser usado para selecionar atributos que obtiveram um ganho de informação médio
Twoing	Alternativa ao índice Gini, melhorando o desempenho quando o atributo de classe tem um domínio grande.
ORT	Critério binário baseado no cálculo do ângulo entre vetores que representam a distribuição de classes nas partições geradas
.Kolmogorov_Sminorv	Critério binário baseado na distância Kolmogorov_Sminorv. Possui uma versão estendida para tratar múltiplas classes e dados faltantes.
AUC[4] *Splitting*	Baseado na área sob a curva ROC[5], referente ao uso de cada um dos atributos descritivos.

Fonte: Adaptado de Silva *et al.*

Todas métricas visam reduzir as impurezas durante a construção da árvore. Silva *et al* (2016), relacionam o conceito de impureza como a variabilidade de classes presentes do atributo-alvo ao particionar um subconjunto. Uma das formas de medição de impureza é a entropia. Eles definem entropia como "grau de desordem num sistema fechado. Numa classificação, a entropia mede o grau de desordem das classes alvos". Witten e Frank (2000) e Provost e Fawcett (2016) detalham a forma de cálculo de obtenção da entropia, esclarecendo a fórmula:

Entropia (p1, p2, .., pn) = -p1*log2 p1 – p2 *$\log_2$ p2 - ...- n*log2 pn (1)

Da Equação 2, o somatório de p é igual a um e cada pi é a relação entre a quantidade de exemplares pertencente da classe-alvo e a do conjunto de dados ao adotar um atributo como critério de divisão de um conjunto de dados.

$$\text{Entropia } (S) = \sum_{i=1}^{c} - pi\ log2\ pi \qquad (2)$$

A partir disso, escolhe-se a dupla atributo e valor de menor entropia como primeiro atributo de quebra do subconjunto, formando ramos de critério de quebra. Este procedimento é realizado em cada sub-árvore recursivamente até formar conjuntos com um único exemplar ou com exemplares pertencentes a mesma classe-alvo.

A utilização de todas as quebras de uma árvore de decisão leva a perda da capacidade de generalização, fenômeno denominado de Sobreajuste (*Overfitting*). Provost e Fawcett (2016, p.113) definem Sobreajuste como *"a capacidade do procedimento de classificação em adaptar-se aos dados, à custa da generalização"*, ou seja, perda da capacidade de aplicação do modelo em outra amostra semelhante.

[4] AUC – Area da Curva ROC
[5] *Receiver Operating Characteristic*

2.1.2. Critério de Generalização do Modelo

O Sobreajuste é um problema inato dos modelos preditivos. Pode-se minimizar seus efeitos, atuando no processo de geração do modelo e na sua avaliação.

A capacidade de generalização do modelo depende dos dados que servirão de entrada durante a sua geração e de validação do modelo. Como estratégia de geração de dados, cabe utilizar a base de dados em treinamento e teste, citado por Silva *et al*:

- Resubstituição, ao utilizar a mesma base de dados para induzir e avaliar o modelo sem distinguir os dados de treinamento e dos de teste. Emprega-se a Resubstituição para verificar a qualidade da codificação do algoritmo e/ou adequação da técnica escolhida;
- *Holdout*, ao utilizar a base de treinamento para induzir o modelo, diferente da base de teste para avaliar o modelo, normalmente, adotando a proporção 70% e 30% da base de dados, como treinamento e teste, respectivamente. Além disso, pode-se empregar o uso dos dados de retenção, dados usados como validação durante o treinamento;
- Validação Cruzada, ao dividir os dados de entrada em K subconjunto em que um subconjunto serve de teste e o K-1 conjuntos de treinamento. Em seguida, escolhe um outro subconjunto para servir de teste e os demais de treinamento até que todos os subconjuntos sejam utilizados como teste;
- *Bootstrap*, semelhante ao *Holdout*, exceto pelo fato dos exemplares já sorteados, serem reposto novamente, na geração das bases de treinamento e de teste.

A forma de validação comumente usada de um modelo preditivo é a matriz de confusão. Provost e Fawcett (2016, p. 189) conceituam matriz de confusão como:

> "Uma matriz de confusão envolvendo n classes é uma matriz n x n com colunas rotuladas como reais e as linhas rotuladas como classe prevista"

Já Silva *et al* (2016) definem a estrutura da matriz de confusão em que as colunas são rotuladas como classe prevista e as linhas, como classes reais, sendo matriz-transposta daquela definida por Provost e Fawcett (2016, p. 190):

> "As linhas são indexadas, seguindo as classes previstas e as colunas são indexadas, seguindo as classes preditas."

Doravante, este artigo adotará a última convenção de estrutura da matriz de confusão. Sendo um classificador binário, a classe-alvo só pode assumir dois valores, positivo ou negativo, quer seja na classe prevista ou real, quer seja na classe predita. Com isto, a matriz de confusão terá duas linhas e duas colunas. Logo, interpreta-se os valores de cada célula conforme Silva *et al* (2016, p. 131):

> "Verdadeiro positivo (VP): classificação correta na classe positiva – o exemplar pertence a classe positiva e o classificador classificou como pertencente a classe positiva.
> Falso positivo (FP): classificação incorreta na classe positiva – o exemplar pertence a classe negativa e o classificador classificou como pertencente a classe positiva.
> Verdadeiro negativo (VN): classificação correta na classe negativa – o exemplar pertence a classe negativa e o classificador classificou como pertencente a classe negativa.

Falso negativo (FN): classificação incorreta na classe negativa – o exemplar pertence a classe positiva e o classificador classificou como pertencente a classe negativa.

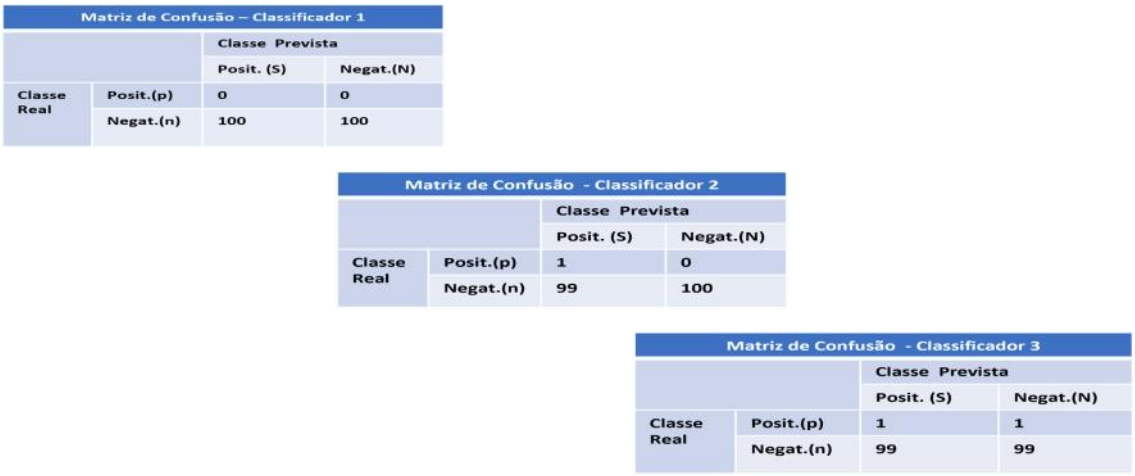
Cabe salientar que a acurácia do classificador é o percentual de acerto, quer seja verdadeiro positivo, quer seja verdadeiro negativo, em relação a quantidade total de exemplares. Ademais, o tipo de erro é um aspecto importante a analisar, entre os quais (Silva *et al* ,2016):

- Sensibilidade ou revocação (do inglês, *recall*) ou taxa de verdadeiros positivos (VP/(VP+FN)) ;
- Especificidade ou taxa de verdadeiros negativos ((VN/ (VN+FP))
- Taxa de falsos positivos (FP/(VN + FP));
- Taxa de falsas descobertas (FP/(VP +FP))
- Preditividade positiva ou precisão (do inglês, precision) ((VP/(VP+FP));
- Preditividade negativa ((VN/(VN+FN));
- *F-score*: relação entre precisão e a revocação (2/ ((1/revocação) + (1 /precisão))

Além da matriz de confusão, pode-se empregar ferramentas gráficas para avaliar o desempenho do classificador, por exemplo, o gráfico bidimensional *Receiver Operating Characteristics* (gráfico ROC), onde a taxa de falso positivo situa-se no eixo x e a taxa de verdadeiro positivo, no eixo y. A linha diagonal dos pontos (0,0) até o ponto (1,1) representa o modelo de comportamento estocástico. O comportamento estocástico ou aleatório significa que a probabilidade do exemplar pertencer a classe positiva é igual a da classe negativa (50%) já que não haveria nenhuma informação sobre os dados neste momento. Logo, o classificador encontrado deve ter desempenho superior ao modelo estocástico, acima dessa diagonal.

O gráfico ROC é construído a partir das diversas gerações da matriz de confusão com a classe positiva e a negativa da classe prevista ou real (classe-base) de acordo com a Figura 2. Em seguida, gera um erro ao trocar o valor do resultado do exemplar de positiva para negativa ou vice-versa. Por exemplo, se o exemplo for verdadeiro positivo, troca-se o resultado para falso positivo, gerando um classificador diferente. Se o exemplo for verdadeiro negativo, troca-se o resultado para falso negativo, gerando um classificador diferente. Este mecanismo é realizado até que todos os exemplares se tornem falso positivo ou falso negativo. Em seguida, traça-se uma curva com as taxas dos falsos positivos e verdadeiro positivo.

Figura 2 Exemplo de uma Matriz de Confusão na Geração Curve ROC



Matriz de Confusão – Classificador 1

		Classe Prevista	
		Posit. (S)	Negat.(N)
Classe Real	Posit.(p)	0	0
	Negat.(n)	100	100

Matriz de Confusão - Classificador 2

		Classe Prevista	
		Posit. (S)	Negat.(N)
Classe Real	Posit.(p)	1	0
	Negat.(n)	99	100

Matriz de Confusão - Classificador 3

		Classe Prevista	
		Posit. (S)	Negat.(N)
Classe Real	Posit.(p)	1	1
	Negat.(n)	99	99

Fonte: Autores.

O gráfico ROC pode não ter boa aceitação perante leigos da área da Ciência de Dados quando desejam comparar diferentes classificadores. Em virtude disso, pode-se calcular a Área do gráfico ROC (AUC), métrica que representa a qualidade do classificador num único valor que varia de 0% a 100%.

Uma forma gráfica de verificação do ponto ótimo, local onde não há Sobreajuste, é interpretar o gráfico complexidade por AUC dos dados de teste e de treinamento. A complexidade da Árvore de Decisão depende da quantidade de nós. O ponto ideal é o momento em que o desempenho da curva dos dados de teste estabiliza ou regride e o dos dados de treinamento continua aumentando. Para a Árvore de Decisão, este valor é chamado de poda ótima, local em que se deve parar de gerar nós da árvore, evitando a perda generalização (Sobreajuste). Este ponto pode ser obtido numericamente através do cálculo do *Complexity Prunning* (*CP*[6]). CP varia entre zero e um. Quanto mais próximo de um, o modelo torna-se menos complexo. Quanto mais próximo de zero, o modelo torna-se mais complexo. A complexidade numa arvore de decisão é medida pelo número de quebras (*nsplit*). A escolha do número de divisões depende do erro encontrado, tendo o erro da validação cruzada (*xerror*) mais comumente utilizado. Uma regra bastante usada é utilizar o *CP* do menor o *xerror*. Caso o *CP* escolhido não produza quebra de árvore (*nsplit*=0), pode-se utilizar o próximo *CP* em que *nsplit* seja diferente zero.

3. Metodologia

Os autores do artigo preferiram seguir a proposta de metodologia de Goldschmidt e Passos (2005). Estes propõem quatros momentos:

- Primeiro momento: Formado pelo Levantamento Inicial ao examinar a base de dados preliminarmente e a Definição dos Objetivos em busca de tarefas de mineração de dados que possam atender a demanda solicitada;

[6] Para definição ver https://cienciaenegocios.com/o-que-e-arvore-de-decisao-decision-tree-linguagem-r/

- Segundo Momento: Planejamento das Atividades, detalhando os planos de ação comas diversas alternativas;
- Terceiro Momento: Plano de Ação, a execução efetiva das tarefas e métodos de KDD
- Quarto Momento: Avaliação dos Resultados ao confrontar os resultados obtidos com as expectativas em relação ao modelo formuladas na etapa de Definição de Objetivos do primeiro momento.

A base de dados utilizada é oriunda da pesquisa[7] realizada no Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), em 2017, sobre consumo de álcool e drogas, relacionamentos, comportamentos de risco e estilo de vida, abrangendo um universo de 19.750 estudantes, obteve-se os primeiros resultados os quais foram divulgados em 2018, no Congresso do IFBA/Campus Salvador, XII Congresso Norte e Nordeste de Pesquisa e Inovação – XII CONNEPI e I Simpósio Brasileiro de Teoria e Ciência das Redes – SimBraRedes. (LIMA-NETO et al, 2018a-c).

Os resultados de cada passo, sugerido pelos autores, serão registrados no Quadro 2.

Quadro 2 - Passos da Metodologia Aplicada

Métodos	**Obs. sobre Parâmetros**	**Resultados**
Carga Registro	· Sem Parametrização	
Transformação dos valores dos atributos nominais em categóricos	· Adotar a idade como 2 categorias: menor ou maior de 18 anos para idade; · Adotar todos os valores como categoria de cada atributos, exceto para a resposta aberta outros, assumindo um outros como um único valor; · Adotar "Ninguém Usa" ou "Alguém Usa" nas questões sobre Antecedentes familiares e Amigos; · Adotar "Usei nos últimos 12 meses" ou "Usei nos últimos 30 dias" ou "Não Usei" nas questões sobre Uso de Substâncias; · Adotar "Sim" ou "Não" nas questões sobre atitudes contravencionais; · Adotar "Sim" ou "não" como classe para a questão "Você já bebeu antes?"	
Seleção de registros	· Todos os registros do IFBA	
Exclusão dos atributos inadequados	Exclusão de atributos que não contribuem na geração do classificador, tais como: identificadores dos respondentes; variáveis contínuas, exceto idade; atributos indicadores de razões do uso ou não de álcool e/ou drogas ilícitas; atributos indicadores de condição econômica,	
Histograma dos principais atributos da árvore completa	· Sem parametrização	
Determinação do Nível de Poda da Árvore	· Menor valor de *error* menor que 1 (*Xerror* do CP) ou informado pelo usuário	
Histograma dos principais atributos de arvore podada	· Poda= CP=Especificada anteriormente	
Geração da Árvore de Decisão Podada	· Mínimo de um elemento nas folhas; · 10 partições para validação cruzada (tenfold cross-validation)	
Partição do BD em treino e teste	· 75% de Treino 25% Teste	
Verificação de Sobre Ajuste		
Determinação do Nível de Poda da Árvore de Decisão da base de Treinamento	· Menor valor de *error* menor que 1 (*Xerror* do CP) ou informado pelo usuário	
Histograma dos principais atributos da base de treinamento	· Poda= CP = Informada	

[7] Submetida a Comitê de Ética em Pesquisa sob o código CAAE: 73745317.7.0000.5031, aprovada conforme o Parecer Circunstanciado nº 2.307.870, em 01 de outubro de 2017.

Geração da Árvore de Decisão de Treinamento Podada	· Mínimo de dois elementos nas folhas; · 10 partições para validação cruzada *(tenfold cross-validation*)	
Predição do classificador através do uso da base de teste e treinamento	· Sem parametrização	
Geração da Matriz de Confusão	· Sem parametrização	
Geração da Curva ROC	· Sem parametrização	

Fonte: Goldschmidt e Passos (2005).

4. Resultados Obtidos

Inicialmente, a base original possuía 619 registros com 153 questões, em que. 14 das 153 questões tiverem seus valores transformados em categóricos conforme parâmetro do método Transformação dos Valores das questões nominais e de acordo com os critérios abaixo:

- Adotar a idade como 2 categorias: menor ou maior de 18 anos para idade;
- Adotar todos os valores como categoria de cada atributos, exceto para a resposta aberta outros, assumindo um outros como um único valor;
- Adotar "Ninguém Usa" ou "Alguém Usa" nas questões sobre Antecedentes familiares e Amigos;
- Adotar "Usei nos últimos 12 meses" ou "Usei nos últimos 30 dias" ou "Não Usei" nas questões sobre Uso de Substâncias;
- Adotar "Sim" ou "Não" nas questões sobre atitudes contravencionais; e
- Adotar "Sim" ou "Não" como classe para a questão "Você já bebeu antes?"

Somente, 256 registros do *Campus* Salvador foram mantidos na base. O especialista da área de saúde sugeriu retirar 94 das 153 questões, restando apenas 59 questões, representadas no Quadro 3. A taxa de base [8] é 83,59% de Bebe_Sim e 16,41% de Bebe_Não.

Quadro 3 - Questões a serem Analisadas.

Cod.	Descrição	Cod.	Descrição
SD1	Idade	SEV38	Tem dificuldades no serviço (seu trabalho é penoso, causa sofrimento)?
SD2	Sexo do entrevistado	SEV39	É incapaz de desempenhar um papel útil em sua vida?
SD4	Modalidade	SEV40	Tem perdido o interesse pelas coisas?
SD6	Turno predominante das aulas	SEV41	Você se sente uma pessoa inútil, sem préstimo?
SD8	Qual sua cor ou raça (segundo o IBGE)?	SEV42	Tem tido ideias de acabar com a vida?
SD9	Qual destas notas (conceitos) você tira com maior frequência?	SEV43	Sente-se cansado(a) o tempo todo?
SD10	Você mora com quem?	SEV44	Tem sensações desagradáveis no estômago?
SD11	Quantas pessoas moram na sua casa?	SEV45	Você se cansa com facilidade?
SD12	Você trabalhou com remuneração ou recebeu bolsa de estudo nos últimos 6 meses?	SEV46	Amigos íntimos achariam se você fumasse um ou mais maços de cigarros por dia
SD13	Quem é o chefe da família? (Em caso de dúvida do entrevistado, eleger aquele que tiver maior instrução).	SEV47	Amigos íntimos achariam se você experimentasse maconha uma ou duas vezes
SD14	Qual é o grau de instrução do chefe da família?	SEV48	Amigos íntimos achariam se você fumasse maconha ocasionalmente
SD16	Com relação à quantidade de comida que há em sua casa você diria que:	SEV49	Amigos íntimos achariam se você fumasse maconha regularmente
SD17	Preferência religiosa	SEV50	Amigos íntimos achariam se você experimentasse crack uma ou duas vezes
SD18	Você pratica a sua religião?	SEV51	Amigos íntimos achariam se você usasse crack ocasionalmente

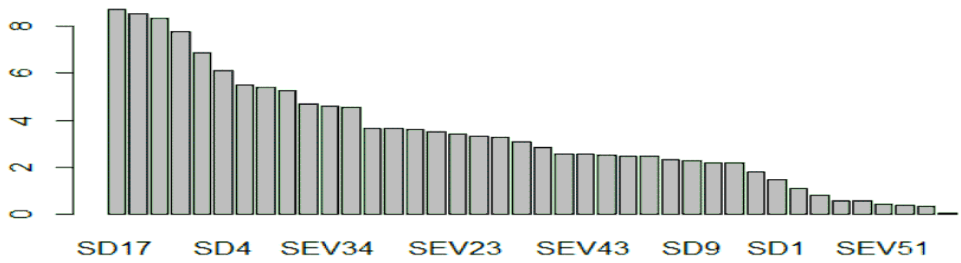
[8] Taxa de base é o percentual de positivos e negativos da base de dados

SEV20	Você está satisfeito com seu peso?	SEV52	Amigos íntimos achariam se você experimentasse cocaína uma ou duas vezes
SEV22	Quantos dias você faltou à escola nos últimos 30 dias?	SEV53	Amigos íntimos achariam se você usasse cocaína de vez em quando
SEV23	O que você faz, em geral, quando falta às aulas?	SEV54	Amigos íntimos achariam se você usasse solvente (loló/lança-perfume, cola) de vez em quando
SEV24	Seus pais ou padrastos vivem (na ausência de pais ou padrastos, considerar o casal responsável pela família).	SEV55	Amigos íntimos achariam se você tomasse um ou dois drinques (bebida alcoólica) quase todo dia
SEV25	Acha que recebe o apoio emocional de que necessita, em qualidade e quantidade, de alguém, amigo (a), familiar, namorado (a), etc?	SEV56	Amigos íntimos achariam se você tomasse cinco ou mais drinques algumas vezes em finais de semana
SEV26	Tem dores de cabeça frequentes?	AFA100	Considerando os últimos 12 meses, algum membro de sua família que mora na mesma casa bebeu a ponto de causar problemas em casa, no trabalho, ou com amigos?
SEV27	Tem falta de apetite?	V132	Você se sentiu ameaçado(a)/humilhado(a) por colegas/alunos de sua escola
SEV28	Dorme mal?	V133	Você já ameaçou/humilhou algum colega ou aluno de sua escola
SEV29	Assusta-se com facilidade?	SEV150	Você realiza ao menos 30 minutos de atividades físicas moderadas ou intensas, de forma contínua ou acumulada, 5 ou mais dias na semana?
SEV31	Sente-se nervoso(a), tenso(a) ou preocupado(a)?	SEV151	Ao menos duas vezes por semana você realiza exercícios que envolvam força e alongamento muscular?
SEV32	Tem má digestão?	SEV152	No seu dia a dia, você caminha ou pedala como meio de transporte e, preferencialmente, usa as escadas ao invés do elevador?
SEV33	Tem dificuldade de pensar com clareza?	BlocoE	ANTECEDENTES FAMILIARES E AMIGOS: Uso de Alguma Droga
SEV34	Tem se sentido triste ultimamente?	BlocoF	USO DE SUBSTÂNCIAS: Uso de Alguma Droga
SEV35	Tem chorado mais do que de costume?	BlocoV134V149	VITIMIZAÇÃO: Foi ou provocou vítima
SEV36	Encontra dificuldades para realizar com satisfação suas atividades diárias?		
SEV37	Tem dificuldades para tomar decisões?	SEV57	Voce já bebeu antes?

Fonte: Autores.

Em seguida, aplicou-se a Árvore Decisão para geração do classificador, obtendo 40 questões, representado no Gráfico 1. As seis primeiras questões com maior capacidade de generalização, nesta ordem, são: SD17, SEV20, SEV24, SEV56, SD8 e SD4. O CP foi gerado. Em seguida, obteve-se zero como valor ótimo do nível máximo de profundidade da árvore automaticamente (*nsplit*). Por isso, o especialista em descoberta de conhecimento em base de dados sugeriu assumir 4 para *nsplit*, o próximo nível máximo da tabela CP. A partir dele, adotou 0,047619 como parâmetro de poda da árvore (CP) de acordo com a Tabela 1.

Gráfico 1 - Atributos com Maiores Níveis de Generalização da Árvore de Decisão Completa



Fonte: Autores.

Tabela 1 - Sugestão de Poda da Árvore

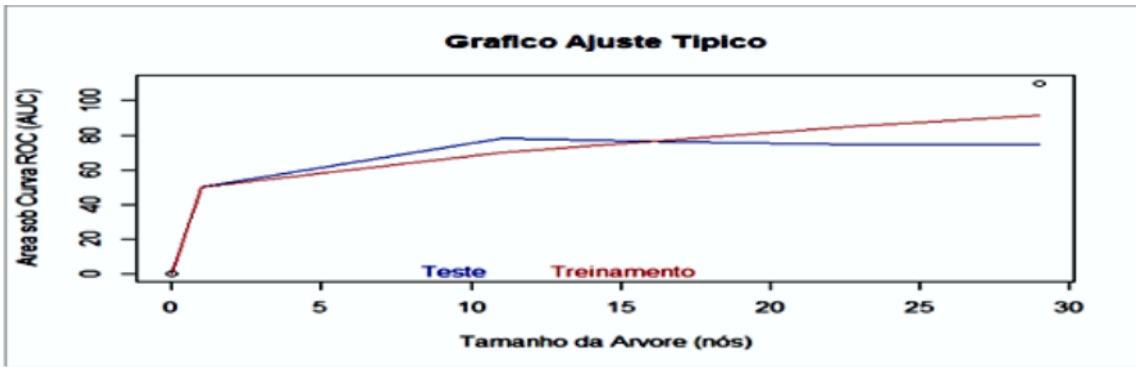
CP(1)	*nsplit(2)*	*rel error(3)*	*Xerror(4)*	*Xstd (5)*
0,0555556	0	1,0000000	1,0000000	0,1410790
0,0476190	4	0,7619048	1,1904762	0,1510252
0,0357143	13	0,3333333	1,1666667	0,1498697
0,0238095	15	0,2619048	1,3333333	0,1574852
0,0158730	19	0,1666667	1,4761905	0,1632027
0,0119048	22	0,1190476	1,5238095	0,1649572
0,0100000	28	0,0476190	1,5476190	0,1658075

Fonte: Autores.

Nota: (1) *Complexity Prunning* *(2)* Número de quebra da árvore (3) Erro Relativo
(4) Erro da Validação Cruzada (5) Desvio Padrão do Erro da Validação Cruzada

O Gráfico 2 sugere 11 nós de tamanho da árvore, que está associado ao CP ótimo de 0,059999 e AUC de 78,14%.

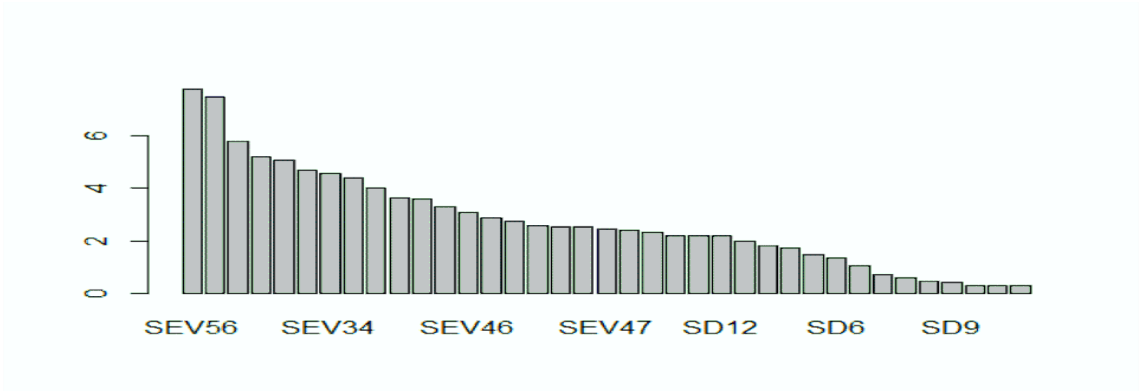
Gráfico 2 - Curva de Ajuste da Árvore de Decisão



Fonte: Autores.

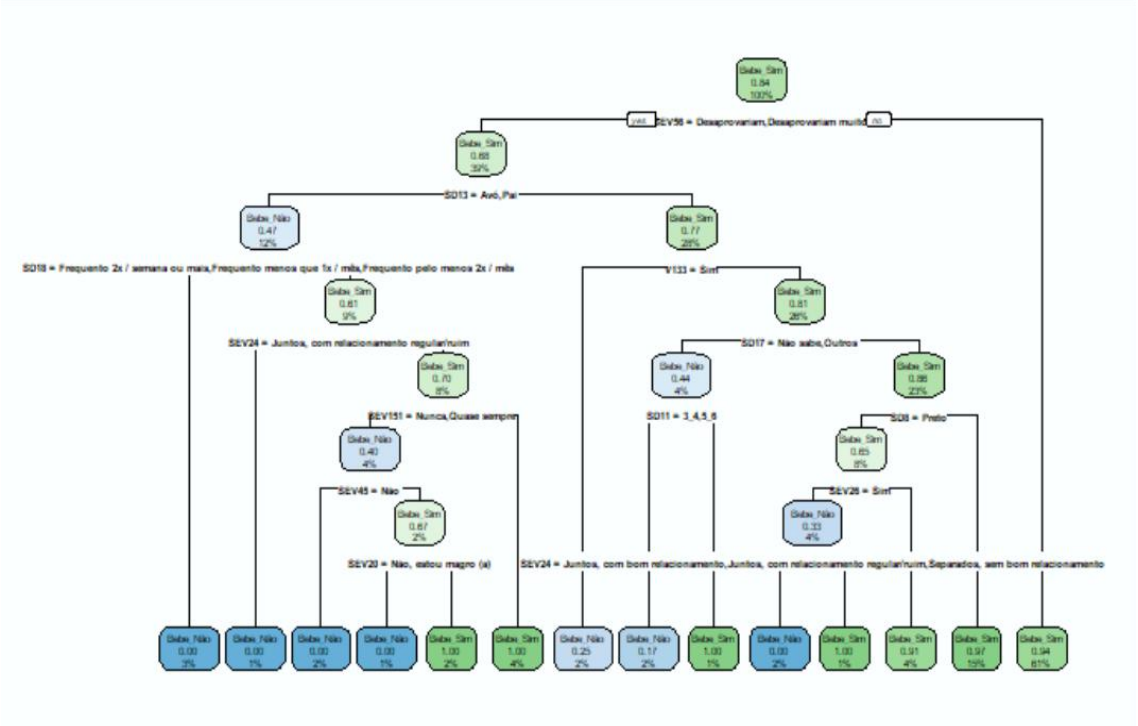
Uma nova Árvore de Decisão foi gerada com parâmetro de poda assumido, obtendo 37 questões, representado no Gráfico 3. As sete primeiras questões com maior capacidade de generalização, nesta ordem, são: SEV56, SEV24, SD17, SEV20, SD11, SEV45 e SEV34. Essas questões não correspondem, diretamente, as questões encontradas da etapa anterior. A árvore de decisão da Figura 3 possui 7 níveis de quebra (*nsplit*) e 14 nós do tipo folha, logo está gerando 14 regras de decisões com até 7 expressões lógicas cada.

Gráfico 3 - Atributos com Maiores Níveis de Generalização da Árvore de Decisão Podada



Fonte: Autores.

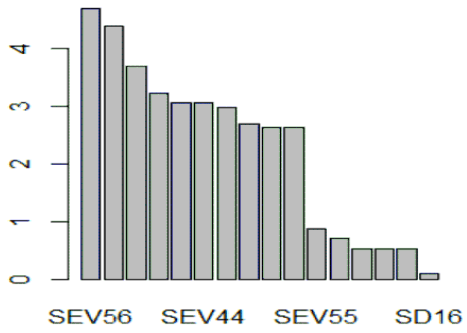
Figura 3 - Árvore de Decisão Podada



Fonte: Autores.

A base original foi dividida, de forma aleatória, numa proporção de 75% e 25%, referente a base de treinamento e de teste, respectivamente. A Árvore de Decisão da base de treinamento foi criada com, no mínimo, dois exemplares nas folhas e 10 partições para validação cruzada (*tenfold cross-validation*), obtendo 16 questões, exibida no Gráfico 4. As quatro primeiras questões com maior capacidade de generalização, nesta ordem, são: SEV56, SEV26, SD4, SEV23. Essas questões não correspondem, diretamente, as questões encontradas da etapa anterior.

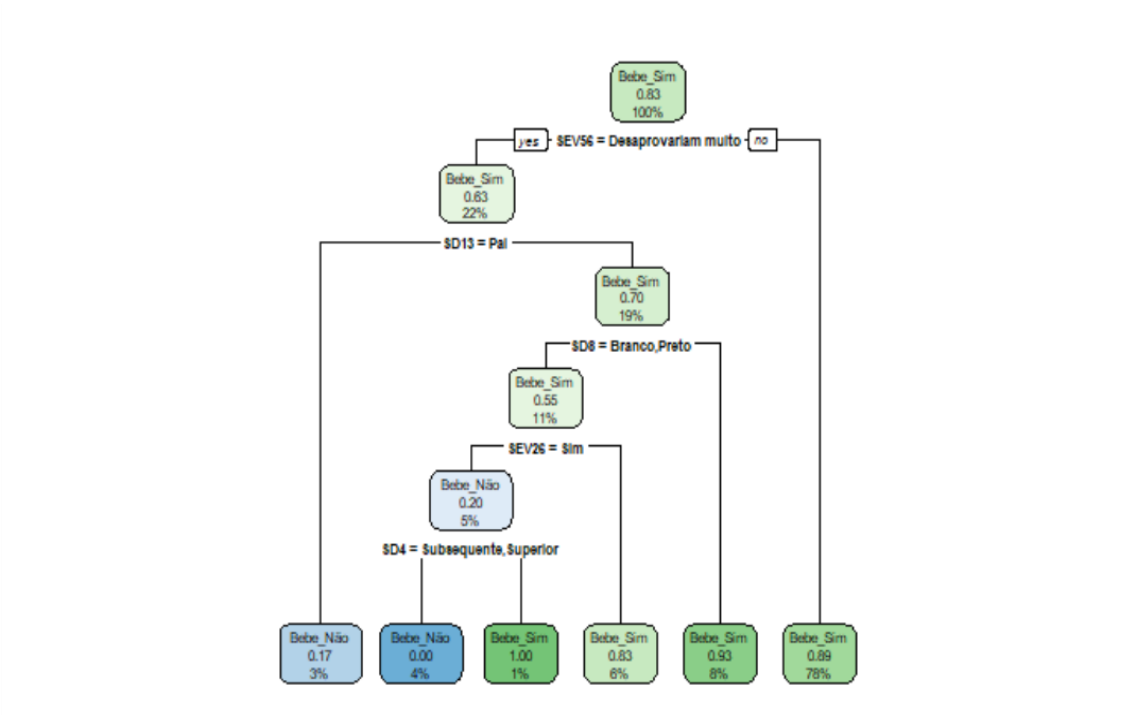
Gráfico 4 - Atributos com Maiores Níveis de Generalização da Árvore de Decisão a partir dos Dados de Treinamento



Fonte: Autores.

O valor de CP foi alterado para 0,059999, valor sugerido. As questões eleitas para a geração da árvore de decisão foram: SD4, SD8, SD13, SEV26 e SEV56. A árvore gerada possui 5 níveis de quebra (*nsplit*) com 6 nós do tipo folha, logo, que podem originar em 6 regras de decisão com no máximo 5 expressões lógicas, representada na Figura 4.

Figura 4 Árvore de Decisão a partir dos Dados de Treinamento



Fonte: Autores.

Nota-se que 0,83% usam álcool do nó do tipo raiz. Considerando que os identificadores e as descrições das questões da Tabela 2, a árvore de decisão produziu as seguintes regras ao analisar da esquerda para a direta.

Tabela 2 - Resultado das Regras de Decisão Produzidas pela Árvore de Decisão a partir dos Dados de Treinamento

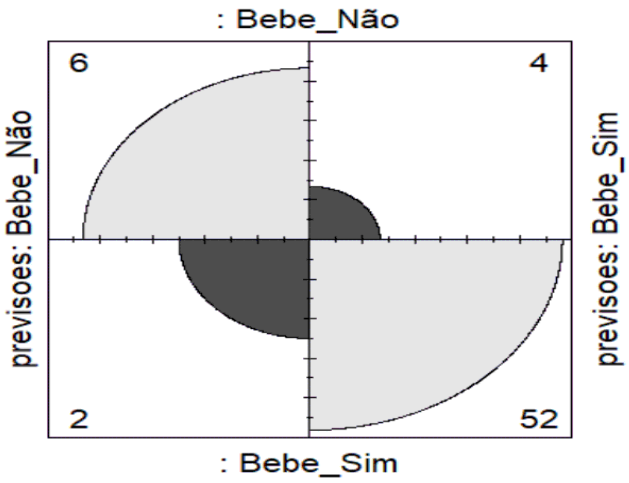
Regra	Prob.[9] Bebe_Sim	Prob, Bebe_Não
Se SEV56 for desaprovaram muito e SD13 for pai	83%	17%
Se SEV56 for desaprovaram muito e SD13 não for pai e SD8 for Branco ou Preto e SD26 for Sim e SD4 for subsequente ou superior	100%	0%
Se SEV56 for desaprovaram muito e SD13 não for pai e SD8 for Branco ou Preto e SD26 for Sim e SD4 não for subsequente e nem superior	100%	0%
Se SEV56 for desaprovaram muito e SD13 não for pai e SD8 for Branco ou Preto e SD26 for Sim	83%	17%
Se SEV56 for desaprovaram muito e SD13 não for pai e SD8 for nem Branco e nem Preto	93%	7%
Se SEV56 não for desaprovaram muito	89%	11%

Fonte: Autores.

[9] Estimativa baseada em Frequência de Probabilidade

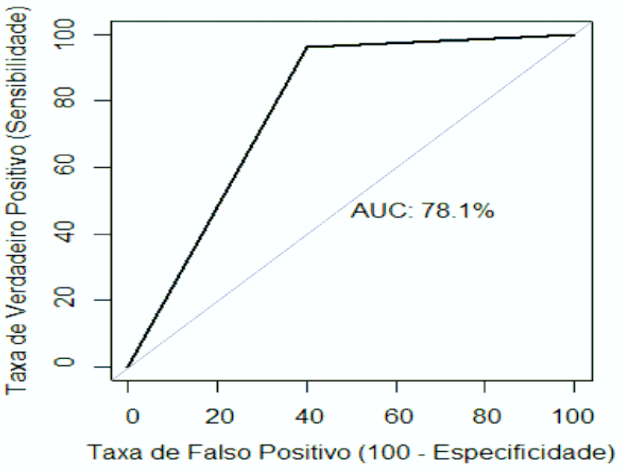
Em seguida, partiu-se para a avaliação do modelo de classificação com a base de teste. O modelo apresentou 6 (9,375%) exemplares como Verdadeiro Negativo referente ao Bebe_Não, 52 (81,250%) exemplares de Verdadeiro Positivo referente ao Bebe_Sim, 2 (3,125%) exemplares de Falso Negativo referente ao Bebe_Sim e 4 (6,25%) exemplares do Falso Positivo referente ao Bebe_Não, representado no Gráfico 5 Logo, sua acurácia foi de 90,62%, superior a 83,59% da classe prevalente da taxa de base. O Gráfico 6 gerado indica desempenho superior ao modelo estocástico, representada pela linha diagonal. Existe diferença aparente entre o Verdadeiro Positivo do Bebe_Sim (81,250%) e o Verdadeiro Negativo do Bebe_Não (9,375%) da base de teste em relação ao Bebe_Sim (83,59%) e ao Bebe_Não (16,41%) da taxa de base, respectivamente. A curva AUC do Grafico 6 atingiu 78,1%. A acurácia e AUC do classificador *Random Forest*[10] com 70 árvores foram 84,38% e 50,0%, respectivamente.

Gráfico 5 - Resultado da Matriz de Confusão



Fonte: Autores.

Gráfico 6 - Curva ROC



Fonte: Autores.

[10] Classificador que utiliza mais de uma arvore para calcular a acurácia e AUC

Por fim, o especialista dobrou a base de dados a partir dela mesma. Agora, a base a ser tratada passou a ter 512 registros. A acurácia do modelo subiu de 90,62% para 92,97% e a área AUC de 78,1% para 88,1%. Se o classificador tivesse sido, também, trocado da Árvore de Decisão para *Random Forest* com 70 árvores, a acurácia e a AUC chegariam a 96,88% e 90,5%, respectivamente.

A redução de 94 das 153 questões, representando cada atributo, contribuiu na simplificação do modelo a partir do conhecimento do especialista em saúde. As 59 questões atributos ainda é uma quantidade alta de questões a serem respondidas por um entrevistado. A aplicação do parâmetro de poda elegeu 16 das 59 questões mais relevantes na tarefa de predição, reduzindo em 72% os números de questões.

5. Considerações Finais

O processo KDD realizado conseguiu gerar um classificador do tipo Árvore de Decisão binária que apresentou resultados significativos em relação a Acurácia e aos tipos de erros Falso Positivo e Falso Negativo. A Acurácia obtida superou a prevalência da taxa de base. As consequências do erro tipo Falso Positivo não compromete nas ações de prevenção, embora o erro do tipo Falso Negativo possa ter uma consequência não desastrosa de imediato.

A utilização de parâmetros de poda simplificou a complexidade do modelo, reduzindo o impacto do Sobreajuste. O modelo simplificado permitiu produzir regras de decisões com facilidade de entendimento e de aplicação nos ambientes informatizados.

Nas condições apresentadas, o classificador *Random Forest* com 70 árvores obteve um desempenho inferior a árvore de decisão. Ao duplicar a quantidade de exemplares, houve pouca melhoria de desempenho da árvore de decisão., surtindo efeito positivo somente no classificador *Random Forest* citado. A decisão de aumentar quantidade de exemplares deve ser avaliada sob a ótica do custo de coleta e o aumento do benefício em termos de Acurácia e de Sensibilidade e Especificidade.

Os autores sugerem realizar uma nova coleta de dados mais recente para verificar a possível degradação ou não do modelo encontrado. A aplicação do modelo em diversos ambientes e momentos diferentes para subsidiar os pesquisadores na elaboração de instrumentos de diagnósticos mais simplificados sobre o uso de álcool.

Em termos de significância prática, esse estudo demonstra que é possível, com até seis perguntas, prever o comportamento de um potencial usuário de álcool, sem haver constrangimento de pergunta direta sobre consumo ou não de álcool.

Referências

BERSON, A.; SMITH, S.J.. Data warehousing, Data Mining & OLAP. p. 333,351, 353. EUA. McGraw Hill, 1997.

LIMA-NETO, J. L. A.; SOUZA, C. R. S.; RIBEIRO, N. M.; FREITAS, M. M.; SANTOS, C.S. Consumo de álcool e drogas por alunos do IFBA: o que os dados nos dizem e o que podemos fazer. (Apresentação) 1º CONGRESSO DO IFBA CAMPUS DE SALVADOR, Junho 2018c. Disponível em: <https://www.researchgate.net/publication/325790744_Presentation_Consumo_de_alcool_e_drogas_por_alunos_do_IFBA_o_que_os_dados_nos_dizem_e_o_que_podemos_fazer>. Acesso em: 6 mar. 2020.

GOLDSCHMIDT, R; PASSOS, E.. **Data Mining: Um Guia Prático**. Conceitos Técnicas, Ferramentas, Orientações e Aplicações. p. 3, 149. Rio de Janeiro: Elsevier Editora, 2005.

WITTEN, I.H.; FRANK, E.. **Data Mining**. Pratical Machine Learning Tools and Techiniques with Java Implementation. p. 3, 93. EUA. Morgan Kaufmann., 2000.

PROVOST, F.; FAWCETT, T.. ***Data Science* para Negócios**. O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico dos Dados. p.14,25,49,113,185,189,190. Rio de Janeiro: Alta Books Editora, 2016.

SILVA, L.; PERES, S.; BOSCARIOLI, C.. **Introdução a Mineração de Dados com aplicações em R.** p. 10,11,79,104,126,130-132. Rio de Janeiro: Elsevier Editora, 2016.

WESTPHAL, C.; BLAXTON, T.. ***Data Mining Solutions***. *Methods and Tools for Solving Real-World Problems.* p. xiv, 123-134,181. Canadá: Wiley, 1998