FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

ALEX RODRIGO KURIBARA

APLICAÇÃO DE CLASSIFICADORES BAYESIANOS E REGRESSÃO LOGÍSTICA NA ANÁLISE DE DESEMPENHO DOS ALUNOS DE GRADUAÇÃO

SÃO PAULO - SP
2015

ALEX RODRIGO KURIBARA

APLICAÇÃO DE CLASSIFICADORES BAYESIANOS E REGRESSÃO LOGÍSTICA NA ANÁLISE DE DESEMPENHO DOS ALUNOS DE GRADUAÇÃO

Projeto de dissertação apresentado à Escola de Administração de Empresas de São Paulo, da Fundação Getulio Vargas, em cumprimento parcial dos requisitos para obtenção do título de Mestre em Administração Empresas.

Área de conhecimento: Sistema da Informação

Orientador: Prof. Dr. Abraham Laredo Sicsú

SÃO PAULO - SP

2015

Kuribara, Alex Rodrigo.
Aplicação de classificadores Bayesianos e regressão logística na análise de desempenho dos alunos de graduação / Alex Rodrigo Kuribara. - 2015.
86 f.

Orientador: Abraham Laredo Sicsú
Dissertação (MPA) - Escola de Administração de Empresas de São Paulo.

1. Estudantes - Brasil - Avaliação de desempenho. 2. Teoria bayesiana de decisão estatística. 3. Análise de regressão logística. 4. Mineração de dados (Computação). I. Sicsú, Abraham Laredo. II. Dissertação (MPA) - Escola de Administração de Empresas de São Paulo. III. Título.

CDU 330.115

ALEX RODRIGO KURIBARA

APLICAÇÃO DE CLASSIFICADORES BAYESIANOS E REGRESSÃO LOGÍSTICA NA ANÁLISE DE DESEMPENHO DOS ALUNOS DE GRADUAÇÃO

Projeto de dissertação apresentado à Escola de Administração de Empresas de São Paulo, da Fundação Getulio Vargas, em cumprimento parcial dos requisitos para obtenção do título de Mestre em Administração Empresas.

Linha de Pesquisa: Sistema da Informação

Data de avaliação: 15 de dezembro de 2015

Banca examinadora:

_____

Prof. Dr. Abraham Laredo Sicsú
FGV-EAESP

_____

Prof. Dr. Alex Aaltonen
Banco Central do Brasil

_____

Prof. Dr. Nelson Lerner Barth
FGV-EAESP

Aos queridos

Victor, Theo e Marisa

Agradecimentos

Ao meu orientador, professor doutor **Abraham Laredo Sicsú** pela paciência e disposição em me ajudar e direcionar nos momentos críticos do desenvolvimento desta dissertação.

Ao **Everton Marques** da Secretaria de Ensino – Curso de Graduação e à **Claudia Castelo Branco** da Coordenadoria de Admissão aos Cursos Regulares da instituição FGV-EAESP por disponibilizarem prontamente os dados utilizados nesta dissertação.

Ao meu filho de 5 anos, **Victor Kenzo**, que teve paciência e maturidade para entender todas as vezes que neguei o seu convite para brincarmos, porque eu estava comprometido em terminar este trabalho.

Agradeço especialmente à minha esposa, **Marisa Akemi**, que me apoiou em todos os momentos desta jornada. Por diversas vezes, pacientemente, ela me ouviu discorrendo sobre o tema, enquanto amamentava o **Theo Hideki**, nosso filho mais novo.

“Education is the most powerful weapon which you can use to change the world.”

Nelson Mandela

RESUMO

Este trabalho minera as informações coletadas no processo de vestibular entre 2009 e 2012 para o curso de graduação de administração de empresas da FGV-EAESP, para estimar classificadores capazes de calcular a probabilidade de um novo aluno ter bom desempenho. O processo de KDD (Knowledge Discovery in Database) desenvolvido por Fayyad et al. (1996a) é a base da metodologia adotada e os classificadores serão estimados utilizando duas ferramentas matemáticas. A primeira é a regressão logística, muito usada por instituições financeiras para avaliar se um cliente será capaz de honrar com seus pagamentos e a segunda é a rede Bayesiana, proveniente do campo de inteligência artificial. Este estudo mostre que os dois modelos possuem o mesmo poder discriminatório, gerando resultados semelhantes. Além disso, as informações que influenciam a probabilidade de o aluno ter bom desempenho são a sua idade no ano de ingresso, a quantidade de vezes que ele prestou vestibular da FGV/EAESP antes de ser aprovado, a região do Brasil de onde é proveniente e as notas das provas de matemática fase 01 e fase 02, inglês, ciências humanas e redação. Aparentemente o grau de formação dos pais e o grau de decisão do aluno em estudar na FGV/EAESP não influenciam nessa probabilidade.

Palavras-chaves: Redes Bayesianas, Regressão logística, KDD, Classificadores Bayesianos, Mineração de dados

ABSTRACT

This dissertation mines a database with information gathered from 2009 to 2012 during the application process to join the business administration course offered by FGV-EAESP. The goal is to develop classifiers which estimate whether a new student will have good performance. The methodology of this dissertation is based on KDD process (Knowledge Discovery in Database) developed by Fayyad et al. (1996a); in addition, the classifiers will be developed by using two theories. The first one is the logistic regression, broadly adopted in financial institutions to assess the potential default of their customers in the credit market. The second one Bayesian networks from artificial intelligence field. The outcomes of this dissertation show that both classifiers have the same discriminant capacity. In addition, the student's age, the number of times she/he applied for FGV/EAESP before joining the school, the region of Brazil she/he comes from and the grades of five exams: Mathematics phase 01 and phase 02, English, Human Science and Essay influence the student performance. However, neither the parents' formal education background nor the student's willingness to join FGV/EAESP impact on such performance.

Keywords: Bayesian Networks, Logistic Regression, KDD, Bayesian Classifiers, Data Mining

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de diagrama para composição de receita da loja de material escolar..... 18
Figura 2 – Grafo direcionado 23
Figura 3 – Exemplo de grafo com ciclo direcionado..... 24
Figura 4 – Exemplo de grafo direcionado para ilustrar nós ascendentes e descendentes..... 24
Figura 5 – Exemplos de grafos em formato de árvore..... 25
Figura 6 – Diagrama para analisar se João será aprovado ou não no vestibular..... 26
Figura 7 – Classificador Naïve Bayes com 4 variáveis previsoras..... 29
Figura 8 – Classificador Naïve Bayes para o exemplo vestibular..... 29
Figura 9 – Classificador Naïve Bayes "alterado" com uma relação de dependência entre as variáveis Desempenho e Prova..... 30
Figura 10 – Classificador TAN com 4 variáveis previsoras 31
Figura 11 – Exemplo de classificador TAN para o exemplo vestibular..... 32
Figura 12 – Classificador BAN com 4 variáveis previsoras 33
Figura 13 – Exemplo de classificador BAN para o exemplo vestibular..... 33
Figura 14 – Exemplo de curva ROC..... 36
Figura 15 – Processo de KDD 37
Figura 16 – Metodologia 40
Figura 17 – Classificador Naïve Bayes com 25variáveis previsoras..... 63

LISTA DE TABELAS

Tabela 1 – Exemplo de base de dados completa 16
Tabela 2 – Definição de potenciais variáveis previsoras e variável alvo para o exemplo do vestibular do João 17
Tabela 3 – Casos possíveis para construção de redes Bayesianas a partir de base de dados... 27
Tabela 4 – Matriz de classificação..... 35
Tabela 5 – Especificidade e sensibilidade na matriz de classificação 35
Tabela 6– Capacidade de discriminação de um modelo de acordo com o valor de ROC 36
Tabela 7 – Número de informações e registros nas bases de dados disponibilizadas para o trabalho..... 48
Tabela 8 – Comparativo entre o número de registro da base de dados e o número de vagas teóricas disponíveis para o curso de administração..... 49
Tabela 9 - Lista de potenciais variáveis previsoras 51
Tabela 10 – Discretização da variável Prova 6 bruta – Redação Fase 02..... 52
Tabela 11 – Fusão das classes da variável Região 52
Tabela 12 – Fusão das classes da variável Questão 06..... 53
Tabela 13 – Análise bivariada para a variável Prova 6 bruta - Redação 54
Tabela 14 - Análise bivariada para a variável Região 54
Tabela 15 – Análise bivariada para a variável Questão 06 54
Tabela 16 – p-valor para o teste de Hosmer-Lemershow 56
Tabela 17 – Índices para avaliar a capacidade de discriminação dos modelos de regressão logística 57
Tabela 18 – Quantidade de variáveis dummies previsores por modelo de regressão logística 57
Tabela 19 – Lista de variáveis previsoras do modelo de regressão logística..... 59
Tabela 20 – Seleção de variáveis baseado no índice de DKL e teste de Chi^2..... 60
Tabela 21 - Comparativo do índice ROC entre classificadores Bayesianos com todas variáveis previsoras e classificadores com as variáveis previsoras selecionadas 61
Tabela 22 – Índice ROC, TAB, TAM e TAT dos 3 classificadores Bayesianos 63
Tabela 23 – Comparativo entre os classificadores Naïve Bayes e o Naïve Bayes Reg.Log.... 64
Tabela 24 – Comparativo entre os índices ROC, TAB, TAM e TAT dos modelos de regressão logística e classificador Naïve Bayes..... 71
Tabela 25 – Grupo de variáveis previsoras presentes nos classificadores..... 72

SUMÁRIO

1 INTRODUÇÃO 12
2 CLASSIFICADORES, REGRESSÃO LOGÍSTICA, CLASSIFICADORES BAYESIANOS E KDD 16
2.1 Classificadores 16
2.1.1 Regressão logística 19
2.1.2 Classificadores Bayesianos 21
2.1.2.1 Noções de probabilidade 21
2.1.2.2 Conceitos de Grafos 23
2.1.2.3 Noções de redes Bayesianas 25
2.1.2.4 Tipos de classificadores Bayesianos 28
2.1.2.4.1 Classificador Bayesiano Naïve Bayes 28
2.1.2.4.2 Classificador Bayesiano Tree augmented Naïve Bayes (TAN) 31
2.1.2.4.3 BN Augmented Naïve Bayes (BAN) 32
2.1.3 Medidas estatísticas para avaliar a capacidade de classificadores 34
2.1.3.1 AUROC – Area Under Receiver Operating Characteristic 34
2.2 Processo de KDD (Knowledge Discovery in Database) 36
3 MATERIAIS E MÉTODOS 39
3.1 Base de dados 39
3.2 Metodologia 39
4 APRESENTAÇÃO DOS RESULTADOS / TRATAMENTO DAS INFORMAÇÕES 46
5 CONCLUSÃO 73
REFERÊNCIAS 77
APÊNDICES 80

1 INTRODUÇÃO

De acordo com o MEC, Ministério da Educação, o sistema brasileiro de educação é composto por quatro etapas. A primeira etapa é a educação infantil para crianças de três a cinco anos, a segunda etapa é o ensino fundamental para crianças de 6 a 14 anos, a terceira etapa é o ensino médio para adolescentes de 15 a 17 anos e a quarta etapa pode ter duas opções para pessoas que concluíram o ensino médio. A primeira opção é o ensino técnico, destinado a qualificar profissionais em diversos setores da economia, como turismo, saúde, tecnologia, contabilidade, entre outros e o seu certificado é equivalente ao ensino médio. A segunda opção é o ensino superior oferecido por universidades, faculdades, institutos superiores e centros de educação tecnológica.

Um aluno que cumpre todas as etapas do sistema de educação brasileiro, ao término do ensino médio, provavelmente passou por um processo de seleção denominado vestibular para ingressar em uma instituição de ensino superior. Em alguns casos, o aluno realizou um curso pré-vestibular para se preparar para as provas do vestibular.

O modelo tradicional do vestibular no Brasil é composto por conjunto de provas podendo ou não ter duas etapas. Quando há duas etapas, na primeira, o candidato realiza um conjunto de provas e de acordo com a sua nota é selecionado ou não para a segunda. Nesta etapa, ele realiza um outro conjunto de provas e pode ser aprovado no vestibular de acordo com o seu desempenho. Em geral, os candidatos se inscrevem em mais de um vestibular, realizando diversas provas para diferentes instituições de ensino superior. Alguns alunos do ensino médio se inscrevem no vestibular como treineiro, candidato que ainda não completou o ensino médio, para conhecer o processo de vestibular.

O processo de vestibular tem sido revisado pelas instituições de ensino e pelo governo federal brasileiro. No início dos anos 2000, o MEC iniciou a aplicação de uma prova chamada ENEM, prova baseada no conteúdo ministrado no ensino médio. Essa prova vem sendo adotada por algumas instituições de ensino superior para selecionar seus alunos.

Em outros lugares do mundo, o critério de seleção para ingressar em uma instituição de ensino superior considera outros fatores além de uma nota de prova de um processo equivalente ao vestibular brasileiro. Nos EUA, os candidatos realizam uma prova padronizada, podendo ser a SAT ou a ACT (MAMLET et VANDEVELDE, 2011). No entanto, além da nota da prova, as instituições podem levar em consideração o engajamento do candidato em atividades extracurriculares, notas do ensino médio, as cartas de motivações explicando o

porquê o candidato quer ingressar na instituição, entrevistas realizadas com o candidato, cartas de recomendação de professores do ensino médio, entre outros fatores.

A FGV-EAESP, faculdade de administração de empresas localizada na cidade de São Paulo no Brasil, oferece curso de administração de empresas e administração pública. Até o vestibular do primeiro semestre de 2011, eram oferecidas 150 vagas para o curso de administração de empresas e 50 vagas para administração pública. Na inscrição, o candidato mencionava em sua ficha a primeira opção de curso e a segunda opção de curso. A partir do segundo semestre de 2011, a FGV-EAESP passou a oferecer 200 vagas para o curso de administração de empresas, 50 vagas para o curso de administração pública e eliminou a opção de cursos. Assim, os candidatos passaram a se inscrever apenas para um curso.

Embora o candidato preencha um questionário no ato da inscrição, as respostas coletadas não são levadas em conta no processo de seleção. A FGV-EAESP utiliza apenas as notas nas provas do vestibular para selecionar os seus futuros alunos. O seu processo de vestibular é composto por duas fases. Na primeira, o candidato realiza provas múltipla escolha de língua portuguesa, literatura e interpretação de textos, matemática, ciências humanas e língua inglesa. Na segunda, o candidato realiza provas dissertativas de matemática e redação. Para ser aprovado para a segunda fase, candidato precisa atingir notas mínimas nas provas da primeira fase e estar classificado entre os n primeiros candidatos, onde n depende do número de vagas disponíveis. A classificação final do vestibular leva em consideração as notas da segunda fase e em alguns anos as notas da primeira fase são consideradas na média final do candidato. Os candidatos são ordenados em ordem decrescente de notas e são convidados para fazer matrículas aqueles que tiveram as maiores notas até que todas as vagas sejam preenchidas.

Todas as informações coletadas durante o processo de vestibular são armazenadas em bancos de dados da faculdade. Na ciência da computação há linhas de pesquisas que focam em maneiras de extrair conhecimento a partir de base de dados. Esses estudos se apoiam no conceito de KDD, *Knowledge Discovery in Database* definido por Fayyad (1996a). No campo da estatística há modelos matemáticos desenvolvidos a partir de bases de dados para tentar discriminar pessoas. Por exemplo, de acordo com Sicsú (2010), as áreas de crédito de instituições financeiras utilizam o modelo de regressão logística para avaliar se o cliente que está solicitando crédito será um bom ou um mau pagador. Uma alternativa à regressão logística pode ser um classificador Bayesiano como visto no trabalho de Karcher (2009).

Este trabalho é um estudo de caso e pode ser de interesse para pessoas que estão iniciando seus estudos em mineração de dados e utilizam classificadores Bayesianos para estimar modelos preditivos. As aplicações podem ser, por exemplo, na área médica para

diagnosticar doenças a partir de sintomas dos pacientes, em marketing para prever hábitos de compra a partir do comportamento dos clientes, em mercado financeiro para prever tendência de preço de ativos a partir de sinais do mercado.

O propósito deste estudo quantitativo é usar o processo de KDD como referência metodológica e estimar um modelo de regressão logística e um classificador Bayesiano a partir das informações capturadas no processo de vestibular. Com isso pretende-se avaliar a possiblidade de identificar alunos que terão bom desempenho no curso de graduação da FGV/EAESP. Nas pesquisas conduzidas durante a dissertação, não foram encontrados estudos na área da educação com enfoque quantitativo para analisar o desempenho de um aluno nos vestibulares. Definição de bom e mau desempenho não foram encontradas e o estudo definiu um critério a ser detalhado nos próximos parágrafos. Assim, esse estudo busca responder a seguinte pergunta de pesquisa:

É possível prever se o aluno de graduação no curso de administração de empresas da FGV-EAESP ingressante entre 2009 e 2012 terá bom desempenho utilizando modelos de regressão logística e classificador Bayesiano nas bases de dados do processo de vestibular da FGV-EAESP?

Para respondê-la, o estudo avaliará as duas hipóteses abaixo:

1. As informações coletadas no processo de vestibular permitem construir um classificador Bayesiano e um modelo de regressão logística;

2. O classificador Bayesiano e o modelo de regressão logística têm poder discriminatório para identificar quais informações indicam que um aluno terá bom desempenho no curso.

O curso de administração de empresas da FGV-EAESP dura no mínimo oito semestres. Nos quatro semestres iniciais, o aluno cursa apenas disciplinas obrigatórias e a partir do quinto semestre ele pode cursar disciplinas optativas e realizar intercâmbios personalizando o seu curso. Embora a estrutura curricular do curso fora alterada oito vezes no período analisado, as disciplinas obrigatórias dos quatro primeiros semestres do curso quase não sofreram alteração. Com isso, acredita-se que utilizar o desempenho dos alunos ao longo dos quatro semestres iniciais garante que eles tiveram desafios acadêmicos semelhantes independentemente o período de ingresso. Neste estudo aluno com bom desempenho será aquele que obteve no máximo uma reprovação por nota. A expressão “reprovação por nota” pode ser denotada pela sigla DP ao longo do texto.

Este trabalho é composto por 5 seções. A seção 02 inicia apresentando o conceito de classificador e regressão logística. Em seguida serão apresentados os conceitos de

probabilidade e teoria de grafos necessários para o entendimento de redes Bayesianas e para enfim detalhar os classificadores Bayesianos. As demonstrações matemáticas e detalhamento de algoritmos não são foco deste trabalho. No final da seção 02, o processo de KDD (*Knowledge Discovery in Database*) é detalhado porque é a base da metodologia adotada. Na seção 03 a metodologia, baseada no processo de KDD, e os critérios utilizados para o tratamento das variáveis e informações são apresentados. A seção 04 apresenta a análise das bases de dados e os resultados dos classificadores Bayesianos e do modelo de regressão logística. A seção 05 apresenta as conclusões extraídas deste trabalho.

2 Classificadores, regressão logística, classificadores Bayesianos e KDD

2.1 Classificadores

De acordo com Friedman et al. (1997), classificação consiste em atribuir uma classe a um elemento em análise a partir de seus atributos. Um classificador determina uma classe a um elemento a partir do valor dos atributos do elemento. As possíveis classes são descritas por uma variável alvo e os atributos por variáveis previsoras.

Esses classificadores podem ser construídos a partir de uma base de dados e na definição dos seus parâmetros é preciso evitar que ocorra *overfitting*[1]. Quando há *overfitting* o classificador avalia com alta acurácia os casos existentes na amostra usada na definição dele. Entretanto, o classificador tem uma acurácia menor quando avalia os demais elementos da população. Isso pode ocorrer devido a especificidade da amostra coletada, sendo mais provável quanto menor a amostra considerada.

Uma base de dados pode ser ou completa ou incompleta dependendo da quantidade de informação faltante para os eventos mapeados. A Tabela 1 possui os elementos da amostra nas linhas e as suas informações nas colunas. Ela é um exemplo de base de dados completa porque há informações em todas as colunas da tabela para cada elemento na linha.

Tabela 1 – Exemplo de base de dados completa

Primeiro nome do candidato	Sexo	Data de nascimento	....	Tipo de instituição de ensino médio
João	Masculino	07/09/1997	....	Pública
Maria	Feminino	12/10/1998	....	Pública
Arthur	Masculino	25/12/1997	....	Privada
Nelson	Masculino	27/07/1998		Privada
Marina	Feminino	21/06/1997		Privada
.	.	.	.	.
.	.	.	.	.
Abraham	Masculino	01/05/1997	....	Privada

Fonte: elaboração própria

Caso a Tabela 1 tivesse algum campo em branco, por exemplo, se não houvesse informação sobre a data de nascimento da Marina, a base seria incompleta. A causa da falta de

[1] *Overfitting* pode ser traduzida para a língua portuguesa como super ajuste.

informação pode ser uma falha na coleta dos dados ou porque a pessoa não quis disponibilizá-la. Segundo Jensen et Nielsen (2007), ao eliminar linhas incompletas da base de dados, o pesquisador pode reduzir a sua amostra ou criar um viés nela. O tratamento de base de dados incompletas não é abordado neste estudo; entretanto, caso o leitor tenha interesse em se aprofundar no tratamento de base de dados incompletas, recomenda-se a leitura de (HECKERMAN, 1995) e (NEOPOLITAN, 2003).

Retornando ao conceito de classificadores, imagine um caso hipotético de João, um jovem que concluiu o ensino médio e está prestando vestibular para a ingressar na FGV/EAESP. Após a realização das provas da segunda-fase, os pais de João ficam na expectativa em saber o resultado. Sabem que seu filho teve um bom desempenho acadêmico no ensino médio e ficou calmo durante as provas do vestibular, embora a concorrência para o curso de administração de empresas fosse alta. Além disso, na primeira fase do vestibular, ele teve um índice de acerto de alto. Com base nisso, os pais de João estão confiantes que ele será aprovado na FGV/EAESP. Analisando o exemplo através da definição de classificador, o elemento em análise é o João, a variável alvo possui dois valores "aprovado no vestibular" e "não aprovado no vestibular". Os atributos levados em consideração são o desempenho acadêmico no ensino médio, o nível de nervosismo durante as provas, a concorrência e as notas do vestibular. O classificador é capaz de determinar se João será aprovado ou não no vestibular utilizando os atributos mencionados, após concluir que eles influenciam na classificação final de João. Um pesquisador, interessado modelar um classificador para esse caso, poderia iniciar o estudo transformando os atributos em potenciais variáveis previsoras e definindo uma variável alvo conforme Tabela 2.

Tabela 2 – Definição de potenciais variáveis previsoras e variável alvo para o exemplo do vestibular do João

Variável	Tipo	Descritivo da variável	Valores
Nervosismo (Ne)	Previsora	Nível de nervosismo do candidato no momento da prova	baixo, médio, alto
Concorrentes (Co)	Previsora	Nível de concorrência por vaga para o curso	baixo, médio, alto
Desempenho (De)	Previsora	Desempenho acadêmico no ensino médio do candidato.	Baixo, médio, alto
Prova (Pr)	Previsora	Nota da prova do vestibular	baixa, média, alta
Aprovação (Ap)	Alvo	Indicação se o candidato foi aprovado no vestibular	sim, não

Fonte: elaboração própria

No campo de aprendizagem de máquina [2], há inúmeras abordagens para a construção de classificadores, por exemplo: árvores de decisão, listas de decisão, redes neurais, grafos de decisão e regras, modelos de regressão, redes Bayesianas, entre outras. Neste estudo, os classificadores são construídos utilizando regressão logística e redes Bayesianas. A primeira é uma ferramenta consagrada para classificação, principalmente na área de análise de créditos em instituições financeiras (SICSÚ, 2010 e SIDDIQI, 2006). A segunda é uma ferramenta baseada em grafos e regras de Bayes utilizada, por exemplo, para identificar diagnóstico de doenças. A sua grande vantagem é representar graficamente a relação, causal ou não (WITTEN et FRANK, 2005), entre as variáveis previsoras e a variável de alvo. Além disso, o estudo de Karcher (2009) mostra que os classificadores Bayesianos produzem resultados tão bons quanto a regressão logística.

Para o mundo corporativo, um diagrama pode ter maior receptividade do que uma fórmula matemática porque é visualmente mais fácil de entender a relação entre as variáveis de um modelo. Por exemplo, considere uma loja de material escolar cujo dono precisa explicar a composição de sua receita mensal para uma pessoa interessada em adquirir o seu estabelecimento. O dono da loja sabe que a receita é composta por venda de cadernos e papéis, lápis e canetas, serviços de fotocópia e encadernação e outros itens. Embora uma fórmula matemática seja capaz de mostrar a composição da receita, como indicado na equação abaixo, provavelmente ele preferirá utilizar um diagrama como a Figura 1 por ser mais visualmente mais fácil de explicar a composição de receita da sua loja.

Receita = (cadernos e papéis) + (lápis e canetas) + (serviços de fotocópia e encadernação) + outros itens

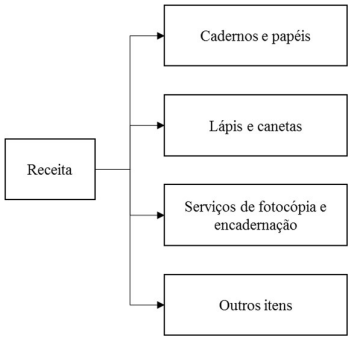

Figura 1 – Exemplo de diagrama para composição de receita da loja de material escolar
Fonte: elaboração própria

[2] A expressão em inglês para aprendizagem de máquina é *machine learning*

2.1.1 Regressão logística

Um modelo de regressão logística estima a probabilidade de um indivíduo pertencer a uma classe a partir de suas características. Nesse modelo as variáveis previsoras são as características do indivíduo e a variável alvo possui as possíveis classes. Esse tipo de classificação é usual em administração de empresas, por exemplo, uma instituição financeira pode utilizar as características do seu cliente para concluir se ele será capaz de pagar ou não o empréstimo. Outro exemplo é uma empresa de telefonia celular que utiliza hábitos de consumo dos usuários e outras características para prever se o seu cliente cancelará ou não o seu contrato de serviço.

Ao desenvolver um modelo de regressão logística, o pesquisador precisa estimar os parâmetros do modelo que melhor descrevam a relação entre a variável alvo e as variáveis previsoras. Quando a variável alvo possui duas classes, utiliza-se a regressão logística binária ou dicotômica. Nos casos de uma variável alvo com mais de duas classes, utiliza-se um modelo de regressão logística multinomial, detalhado em (HOSMER et al., 2013).

Neste estudo, como a variável alvo possui duas classes: "Bom desempenho" e "Mau desempenho", sendo convertidas para 1 e 0 respectivamente, será utilizado o modelo de regressão logística binária.

Se o modelo de regressão logística tivesse apenas uma variável previsora para estimar a probabilidade de um aluno ter bom desempenho, ele seria descrito pela equação 01. Nessa equação, β_0 e β_1 são os parâmetros a serem estimados, X é a variável previsora e $\pi(X)$ a variável alvo.

$$P(Y = bom\ desempenho|X) = \pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1+ e^{\beta_0+\beta_1 X}}$$

(Equação 01)

A transformada logit pode ser utilizada para representar a mesma equação, conforme apresentado na equação 02.

$$z(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = \beta_0 + \beta_1 X$$

(Equação 02)

A razão $\frac{\pi(X)}{1-\pi(X)}$ é a razão de chances do aluno ter bom desempenho, ln é o logaritmo na base e (e=2,71828) e o parâmetro β_1 indica a variação de z(X) para variação unitária da variável previsora X.

Se o modelo fosse composto por n variáveis previsoras, X_1, X_2,..., X_n, ele passaria a ser descrito pela equação 03.

$$P(Y=\text{bom desempenho}|\ X_1, X_2,..., X_n) = \pi(X_1, X_2, \dots, X_n) = \frac{e^{\beta_0+\beta_1.X_1+\cdots+\beta_n.X_n}}{1+e^{\beta_0+\beta_1.X_1+\cdots+\beta_n.X_n}}$$

(Equação 03)

A equação 03 reescrita utilizando a transformada logit, é dada pela equação 04.

$$z(X_1, \dots, X_n) = \ln\left[\frac{\pi(X_1,\dots,X_n)}{1-\pi(X_1,\dots,X_n)}\right] = \beta_0 + \beta_1.X_1 + \cdots + \beta_n.X_n$$

(Equação 04)

No exemplo do vestibular, o modelo de regressão logística, equação 05, seria determinado para o conjunto de potenciais variáveis previsoras {Desempenho, Nervosismo, Concorrentes, Prova} e a variável alvo seria Aprovação. Ela assume valor igual a 1 se o João for aprovado e valor 0 se não for aprovado.

$$\ln\left[\frac{P(Aprovação = 1)}{1 - P(Aprovação = 1)}\right] = \beta_0 + \beta_1\, Desempenho + \beta_2 \text{Nervosismo} + \beta_3 \text{Concorrentes} + \beta_4 \text{Prova}$$

(Equação 05)

Os parâmetros do modelo de regressão logística são estimados através do método de máxima verossimilhança e o seu cálculo pode ser feito por meio de softwares estatísticos como o SPSS, SAS e Minitab. Após estima-los, é preciso fazer um teste de significância do modelo e uma análise de aderência dele em relação a amostra. O teste de significância avalia se as variáveis previsoras se relacionam com a variável alvo. Para isso, realiza-se um teste de hipótese para os parâmetros do modelo, onde H_0: $\beta_1 = \beta_2 \dots \beta_n = 0$ e H_1: pelo menos um $\beta_i \neq 0$, i = 1,....,n.

A análise de aderência avalia se o resultado do modelo de regressão logística é coerente com a amostra e o teste comumente utilizado é o Hosmer-Lemeshow, detalhado em

(HOSMER et al., 2013). Para essa análise, o teste de hipótese é H_0 = o modelo obtido se ajusta a amostra e H_1 = modelo não se ajusta a amostra.

Após analisar o modelo estimado, o pesquisador pode avaliar o quanto uma variável previsora influencia na probabilidade de o elemento estudado pertencer a uma classe da variável alvo. No exemplo do vestibular, o pesquisador pode analisar a influência da variável Desempenho na probabilidade de João ser aprovado no vestibular. O indicador denominado razão de chances[3] mostra o quanto a probabilidade aumenta ou diminui devido a um aumento unitário da variável previsora. Para a variável Desempenho, a razão de chances é dada pela equação 06.

$$\text{razão de chances} = \frac{\pi(Aprovação=1)}{\pi(Aprovação=0)} = e^{\beta_1}$$

(Equação 06)

2.1.2 Classificadores Bayesianos

Classificadores Bayesianos é uma aplicação de Redes Bayesianas, conforme apresentado nos trabalhos de Lengley et al. (1992), Friedman et al. (1997), Cheng (1999) e Jensen et Nielsen (2007). Antes de detalhar os tipos de classificadores Bayesianos utilizados neste estudo de caso, serão apresentados os conceitos de probabilidades, grafos e Redes Bayesianas.

2.1.2.1 Noções de probabilidade

Em classificadores Bayesianos, os conceitos de probabilidades condicionais são utilizados para determinar os parâmetros das variáveis previsoras e atualiza-los. Para ilustrar os conceitos será utilizado o caso hipotético do candidato João, que está prestando vestibular para ingressar na FGV/EAESP.

Enquanto o resultado do vestibular não é divulgado, os pais de João, muito ansiosos para saber o resultado, constantemente conversam sobre o assunto e tentam estimar a

[3] razão de chances é a tradução da expressão utilizada em inglês, odds ratio

probabilidade de ele ser aprovado. Inicialmente, acreditando que o seu filho é o aluno mais inteligente do Brasil e com base nas suas histórias de quando prestaram vestibular, estimam que a probabilidade de João ser aprovado é altíssima. No entanto, após momentos de euforia, notam que estão com um alto viés de otimismo e após pesquisas na internet, descobrem que a probabilidade de um candidato ser aprovado no vestibular é influenciada pelo seu desempenho acadêmico no ensino médio. Após essa descoberta eles se indagam: "Seria possível saber a probabilidade de ser aprovado, sabendo que ele teve um bom desempenho acadêmico no ensino médio?".

O teorema de Bayes é a relação matemática capaz de responder à pergunta. Como exemplo didático, utilizando as variáveis da Tabela 2, assuma a variável alvo Aprovação, como a probabilidade de João ser aprovado no vestibular e Desempenho, como o seu bom desempenho no ensino médio. Para Ap = sim e De = bom, o teorema de Bayes estabelece a equação 07.

$$P(Ap = sim|De = bom) = \frac{P(De = bom|Ap = sim)}{P(De = bom)}.P(Ap = sim)$$

(Equação 07)

A probabilidade inicial determinada pelos pais de João, P(Ap=sim), é denominada probabilidade a *priori* e a nova probabilidade de Ap sabendo que ocorreu o evento De, P(Ap=sim| De=bom) é denominada probabilidade a *posteriori*. Os valores de P(De=bom|Ap=sim) e P(De=bom) podem ser determinados a partir de um histórico de informação.

O nível de concorrência pode afetar a probabilidade de aprovação do João no vestibular, porém a nota de um outro candidato não influencia a probabilidade de João tirar uma nota 10. Quando um evento não influencia a probabilidade de ocorrência de outro evento diz-se que são independentes. Ao denominar nota de João igual a 10 como evento A e nota do outro candidato igual a 10 como evento B, pode-se escrever a relação abaixo:

$$P(A|B) = P(A)$$

A probabilidade de os dois alunos tirarem nota 10, pode ser dada pela relação da equação 08.

$$P(A \cap B) = P(A|B).P(B) = P(A).P(B)$$

(Equação 08)

Pode ser demonstrado que o conceito de independência é simétrico, ou seja, a probabilidade da nota do outro candidato ser 10, independe de João ter uma nota 10 na prova.

A independência também ocorre com eventos relacionados. Dois eventos, A e B, são condicionalmente independentes de um terceiro, C, e a equação 09 é verdadeira.

$$P(A \cap B|C) = P(A|C).P(B|C)$$

(Equação 09)

2.1.2.2 Conceitos de Grafos

Os digramas dos classificadores Bayesianos são representados por meio de grafos e nesta seção são definidos alguns conceitos essenciais para o entendimento de classificadores Bayesianos.

Segundo West (2010), grafos direcionados são diagramas formados por arcos e nós, podendo ser representados por G{N,A}, onde N é o conjunto de nós $\{N_1, N_2....N_k\}$ e A o conjunto de arcos $\{A_1, A_2,....A_m\}$. Cada arco interliga dois nós indicando qual é o nó de origem e o nó de destino. Em classificadores Bayesianos, os nós representam variáveis e os arcos as relações entre elas. Neste estudo, nós e variáveis serão expressões intercambiáveis, assim como grafos e redes.

Nos textos em inglês, quando o valor de um nó é conhecido, diz-se que ele está "instantiated", e a expressão encontrada nos textos em língua portuguesa foi instanciado(a). Assim, a expressão "nó instanciado" é utilizada para dizer que o valor do nó é conhecido.

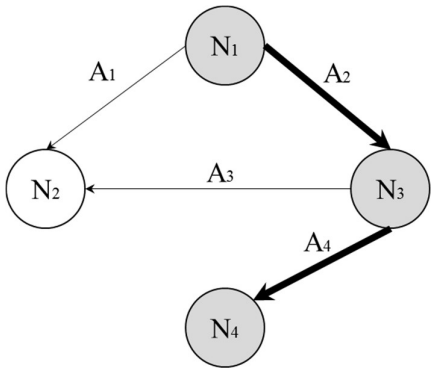

Figura 2 – Grafo direcionado
Fonte: elaboração própria

Denomina-se caminho, o conjunto de arcos que interligam um grupo de nós. Na Figura 2, há um caminho formado por A_2 e A_4 interligando os nós N_1, N_3 e N_4. Se o grafo tivesse a direção do arco A_1 alterada de (N_1,N_2) para (N_2, N_1), conforme Figura 3, o caminho formado por A_2, A_3 e A_1 levaria o nó N_1 a ele mesmo. Quando isso ocorre Neapolitan (2003), define o caminho como um **ciclo direcionado**.

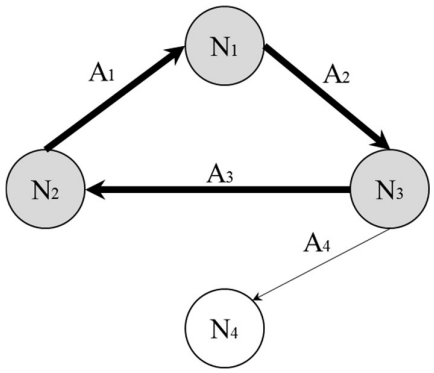

Figura 3 – Exemplo de grafo com ciclo direcionado
Fonte: elaboração própria

Em grafos, os nós possuem diferentes denominações de acordo com a relação entre eles. Em um grafo direcionado, o nó de origem do arco é denominado nó ascendente ou pai, já o nó de destino do arco é denominado nó descendente ou nó filho. A Figura 4 é um grafo formado por cinco nós {N_1, N_2, N_3, N_4, N_5} e cinco arcos {A_1, A_2, A_3, A_4, A_5}. O nó N_1 é o nó ascendente ou pai dos nós N_3 e N_4. O nó N_5 é o descendente ou filho de N_3 e N_4.

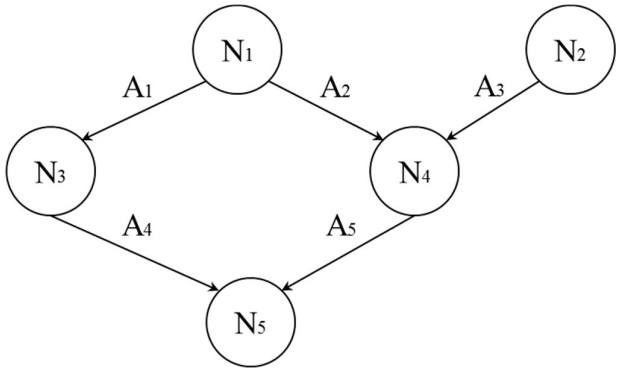

Figura 4 – Exemplo de grafo direcionado para ilustrar nós ascendentes e descendentes
Fonte: elaboração própria

Árvore é um tipo de grafo utilizado em modelos de tomada de decisão e em classificadores Bayesianos. Uma árvore é um grafo sem ciclos e com pelo menos um caminho

interligando dois nós quaisquer pertencentes à árvore, conforme ilustrado na Figura 5. Note que em ambos grafos não há ciclos e há caminhos interligando dois nós, por exemplo, na figura à esquerda há um caminho interligando os nós N_1 e N_6 e um outro interligando N_2 e N_5. Em nenhum dos casos há nós sem ligação, ou seja, nós isolados dos outros que formam o grafo.

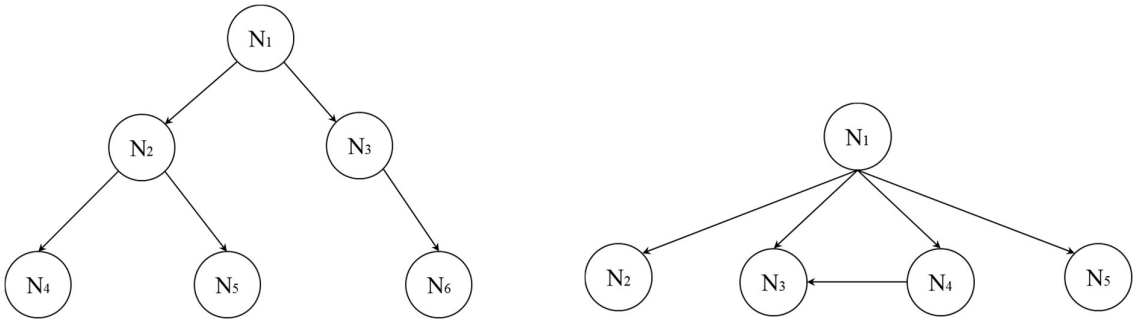

Figura 5 – Exemplos de grafos em formato de árvore
Fonte: elaboração própria

2.1.2.3 Noções de redes Bayesianas

Após apresentar os conceitos de probabilidade e de grafos, é a vez de definir redes Bayesianas e em seguida detalhar os tipos de classificadores Bayesianos.

Uma rede Bayesiana é um grafo direcionado acíclico, cujos nós representam as variáveis e os arcos as relações entre elas. Cada variável possui um conjunto de valores finitos, uma distribuição de probabilidade e um conjunto de nós pais. Além disso, as variáveis são condicionalmente independentes de todas que não são seus pais e não são seus descendentes. Uma rede Bayesiana pode ser utilizada para estimar a probabilidade de ocorrência de um evento a partir de outros eventos relacionados.

Para um conjunto de variáveis, $X=\{X_1, X_2,...,X_n\}$, que formam uma rede Bayesiana, B, e cada variável X_n possui m nós pais, $Pa_{X_n} = \{Pa_1, \ldots., Pa_m\}$, a distribuição de probabilidade conjunta da rede B para um conjunto de valores das variáveis, $x =\{X_1 = x_1, X_2 = x_2,...,X_n = x_n\}$ é dada pela equação 10.

$$P_B(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{Pa}_{X_i})$$

(Equação 10)

No caso do vestibular do João, uma rede Bayesiana possível é composta por cinco nós, representando as variáveis descritas na Tabela 2 e cinco arcos {A_1, A_2, A_3, A_4 e A_5} estabelecendo a relação entre as variáveis. Se os pais de João conhecessem redes Bayesianas, poderiam construir uma conforme Figura 6. Nessa rede, o grau de concorrência no vestibular da FGV/EAESP pode influenciar no nível de nervosismo de João no momento da prova. O nível de nervosismo de João durante a prova e seu desempenho acadêmico no ensino médio podem influenciar a sua nota nas provas do vestibular. A chance de ele ser aprovado depende do grau de concorrência no vestibular e de suas notas nas provas.

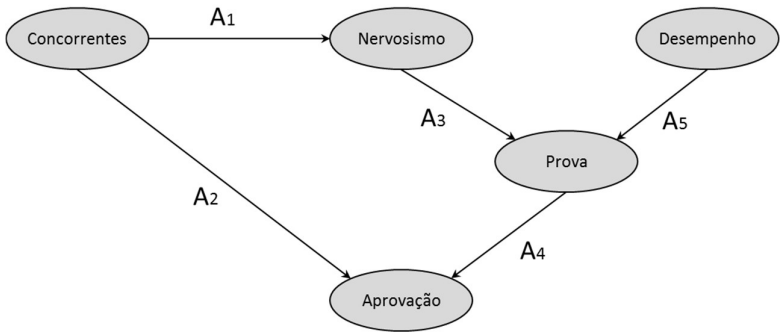

Figura 6 – Diagrama para analisar se João será aprovado ou não no vestibular
Fonte: elaboração própria

A definição de uma rede Bayesiana pode ser feita por três métodos. O primeiro é através do conhecimento de especialistas no assunto em análise. Por exemplo, especialistas da área de educação poderiam contribuir com os pais de João na construção da rede Bayesiana da Figura 6. De acordo com Heckerman (1995), quando a estrutura da rede e os seus parâmetros são definidos a partir do conhecimento de especialistas, diz-se que as probabilidades são Bayesianas. No entanto, em alguns casos o assunto pode ter sido pouco explorado e não haver especialistas, ou possuir um grande número de variáveis envolvidas. Consequentemente a definição da estrutura da rede e dos parâmetros das variáveis se tornam complexas.

O segundo método utiliza base de dados e algoritmos de mineração de dados. A partir de uma base de dados contendo uma amostra com os valores das variáveis em análise, esses algoritmos são capazes de determinar uma rede Bayesiana com as relações entre as variáveis contidas na base de dados e os respectivos parâmetros. Nesse caso, os parâmetros são definidos a partir de frequências relativas, ou seja, pelo método clássico estatístico.

O terceiro método utiliza o conhecimento de especialistas e também extrai conhecimento de bases de dados, ou seja, combina os dois métodos anteriores. Nesse caso, os

especialistas ajustam as relações entre as variáveis e os parâmetros extraídos das bases de dados. Como o interesse deste estudo é extrair conhecimento de base de dados, será utilizado o segundo método.

Um pesquisador utilizando o segundo método para determinar uma rede de Bayes pode se deparar com quatro casos, conforme Tabela 3.

Tabela 3 – Casos possíveis para construção de redes Bayesianas a partir de base de dados

		Base de Dados	
		Completa	Incompleta
Estrutura da Rede	Conhecida	Caso 01	Caso 03
	Não conhecida	Caso 02	Caso 04

Fonte: elaboração própria

No caso 01 a base de dados é completa e a estrutura da rede é conhecida, ou seja, a relação entre as variáveis é conhecida. Nesse caso, o pesquisador deve se preocupar em determinar os parâmetros da rede, como acontece com o classificador Naïve Bayes definido na seção 2.1.2.4.1.

No caso 02, a base de dados é completa, porém a estrutura é desconhecida. A definição da estrutura de uma rede consiste em determinar aquela que melhor descreve a relação entre as variáveis. Uma maneira de definir uma rede é construir um grafo com todas as ligações possíveis entre as variáveis. No entanto, isso pode criar estruturas complexas para o processamento e *overfitting* na base de dados. A alternativa comumente adotada é utilizar algoritmos que determinam uma estrutura de rede levando em consideração a complexidade da rede e a acurácia dela. A estrutura de rede escolhida tem o melhor balanceamento entre acurácia e complexidade.

Os casos 03 e 04 possuem bases de dados incompletas. Quando isso ocorre é preciso tratar a ausência de informações e uma das alternativas é utilizar o algoritmo EM *(Expectation Maximization)* detalhado em Heckerman (1995). Após o tratamento da base de dados, o caso 03 se torna análogo ao caso 01 e o caso 04 análogo ao caso 02.

Após definir a estrutura da rede, os parâmetros são determinados contando os casos na base de dados que se enquadram na relação de dependência definida pela rede.

2.1.2.4 Tipos de classificadores Bayesianos

Um classificador Bayesiano pode apresentar diferentes configurações e neste trabalho serão utilizados três tipos: Naïve Bayes (LENGLEY et al, 1992), o Tree Augmented Naïve Bayes, TAN, (FRIEDMAN et al.,1997) e o Augmented Naïve Bayes, BAN, (FRIEDMAN et al., 1997 e CHENG, 1999). Assim como uma rede de Bayes, o classificador Bayesiano pode ter estrutura conhecida como o Naïve Bayes ou uma estrutura desconhecida como o BAN. A característica em comum desses três classificadores é ter a variável alvo com pai de todas as previsoras e as variáveis previsoras podem ou não ser condicionalmente independentes entre elas, dependendo do tipo de classificador.

Para os classificadores Naïve Bayes, TAN e BAN formados por uma variável alvo, C, com dois valores potenciais c_1 e c_2, e um conjunto de variáveis previsoras, N_1, N_2,...., N_n, cujos valores são representados por letra minúscula do nome das variáveis, pode-se calcular a probabilidade de C assumir valor igual a c_1 pela equação 11.

$$\mathrm{P}(c_1|n_1, n_2, n_3 \dots n_n) = \mathrm{P}(c_1).\frac{\prod_{i=1}^{n} P(n_i|\mathrm{Pa}_{n_i})}{P(n_1, n_2, n_3, \dots, n_n, c_1) + P(n_1, n_2, n_3, \dots., n_n, c_2)}$$

(Equação 11)

Dependendo da estrutura do classificador Bayesiano, o conjunto de pais para cada variável se altera, conforme detalhado nas seções a seguir.

2.1.2.4.1 Classificador Bayesiano Naïve Bayes

O classificador Naïve Bayes, Figura 7, é uma rede de Bayes com uma estrutura definida. A variável alvo é o nó pai de todas as variáveis previsoras e a principal premissa é assumir que todas as variáveis previsoras são condicionalmente independentes entre si dado o valor da variável alvo. Em geral, quando as variáveis previsoras não são fortemente correlacionadas, o classificador Naïve Bayes possui um desempenho melhor (LENGLEY et al, 1992) do que outros classificadores Bayesianos com estruturas de rede mais complexas.

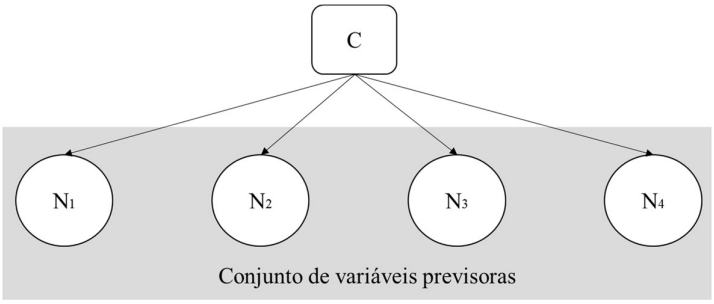

Figura 7 – Classificador Naïve Bayes com 4 variáveis previsoras
Fonte: elaboração própria

De acordo com Friedman et al. (1997), o classificador Naïve Bayes utiliza um método de contagens de casos na amostra para determinar os parâmetros das variáveis previsoras, conhecendo o valor da variável alvo. Para o classificador da Figura 7, a classificação é definida aplicando a regra de Bayes para calcular a probabilidade a *posteriori* de a variável alvo C ser c_1, sabendo os valores das variáveis previsoras N ={n_1, n_2,. n_3, n_4}, conforme equação 12.

$$\mathrm{P}(c_1|n_1, n_2, n_3, n_4) = \mathrm{P}(c_1).\frac{P(n_1|c_1).P(n_2|c_1).P(n_3|c_1).P(n_4|c_1)}{P(n_1, n_2, n_3, n_4, c_1) + P(n_1, n_2, n_3, n_4, c_2)}$$

(Equação 12)

Para o exemplo do vestibular, a variável alvo do classificador Naïve Bayes, Figura 8, é a Aprovação e as variáveis previsoras são Concorrentes, Provas, Nervosismo e Desempenho.

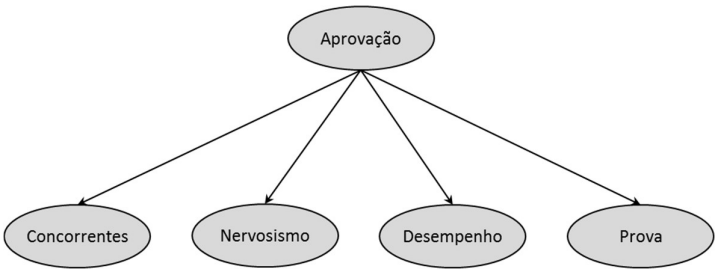

Figura 8 – Classificador Naïve Bayes para o exemplo vestibular
Fonte: elaboração própria

Voltando ao caso do João, seus pais sabiam que ele teve um alto desempenho acadêmico no ensino médio, obteve notas média nas provas, o seu nível de nervosismo era baixo durante as provas e a concorrência no vestibular era alta. O cálculo da probabilidade de o João ser aprovado no vestibular é dado pela Equação 13. Para simplificar a notação será adotado Ap para Aprovação, De para Desempenho, Pr para Prova, Co para Concorrente e Ne

para Nervosismo. Quando a variável Ap assume valor de aprovado, a notação utilizada será de Ap= sim.

$$\mathrm{P}(\mathrm{Ap}=\mathrm{sim}|\mathrm{De}=\mathrm{alto},\mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta}) =$$
$$= \frac{\mathrm{P}(\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{De}=\mathrm{alto}|\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{Pr}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{Ne}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{Co}=\ \mathrm{alta}|\mathrm{Ap}=\mathrm{sim})}{\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{sim})+\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{Não})}$$

(Equação 13)

Se as variáveis previsoras desse exemplo não fossem condicionalmente independentes entre elas, o classificador deixaria de ser um Naïve Bayes e a estrutura da rede deveria ser alterada acrescentando arcos para indicar a condição de dependência entre as variáveis previsoras. Por exemplo, considere que a variável Desempenho influencia Prova, nesse caso é adicionado um arco interligando-as, conforme Figura 9 e o classificador passa a ser um Tree Augmented Naïve Bayes, detalhado na próxima seção. Além disso, a equação 13 deveria refletir essa nova dependência, com a alteração do fator P(Pr=médio|Ap=sim) para P(Pr=médio|Ap=sim, De=médio), como mostra a equação 14.

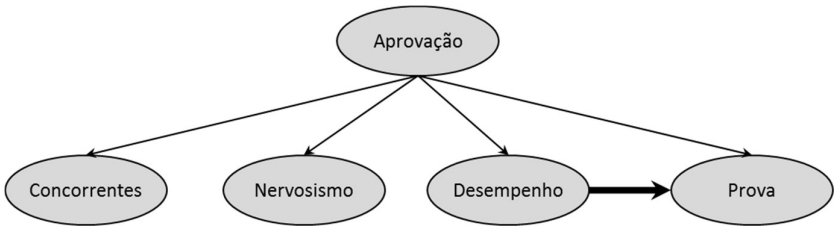

Figura 9 – Classificador Naïve Bayes "alterado" com uma relação de dependência entre as variáveis Desempenho e Prova
Fonte: elaboração própria

$$\mathrm{P}(\mathrm{Ap}=\mathrm{sim}|\mathrm{De}=\mathrm{alto},\mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta}) =$$
$$= \frac{\mathrm{P}(\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{De}=\mathrm{alto}|\mathrm{Ap}=\mathrm{sim}).\mathbf{P}(\mathbf{Pr}=\ \mathbf{médio}|\mathbf{Ap}=\mathbf{sim},\mathbf{De}=\mathbf{médio}).\mathrm{P}(\mathrm{Ne}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{Co}=\ \mathrm{alta}|\mathrm{Ap}=\mathrm{sim})}{\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{sim})+\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{Não})}$$

(Equação 14)

2.1.2.4.2 Classificador Bayesiano Tree augmented Naïve Bayes (TAN)

O classificador Tree Augmented Naïve Bayes (TAN) propõe melhorar o modelo Naïve Bayes. Como no classificador Naïve Bayes, a variável alvo é pai de todas as variáveis previsoras. No entanto, a premissa que todas as variáveis previsoras são condicionalmente independentes entre elas dado o valor da variável alvo deixa de existir.

A estrutura da rede fica restrita a grafos em formato de árvores e os pais das variáveis previsoras são a variável alvo e no máximo uma outra variável previsora. O algoritmo para definição de um classificador Bayesiano TAN é baseado no algoritmo de Chow et Liu (1968) que adiciona à estrutura arcos entre duas variáveis, X e Y, baseado no quanto essa nova conexão vai reduzir o nível de incerteza do valor de X sabendo o valor de Y.

A Figura 10 ilustra um classificador Bayesiano TAN com quatro variáveis previsoras. Nesse caso, o algoritmo adicionou um arco entre N_2 e N_1 e um outro entre N_3 e N_4. Com isso, o conjunto de pais das variáveis previsoras são $P_{N1} = \{C, N_2\}$, $P_{N2} = \{C\}$, $P_{N3} = \{C\}$ e $P_{N4} = \{C, N_3\}$. Assim como no classificador Naïve Bayes, os parâmetros das variáveis são determinados por um método de contagem de casos na amostra utilizada para definir o classificador.

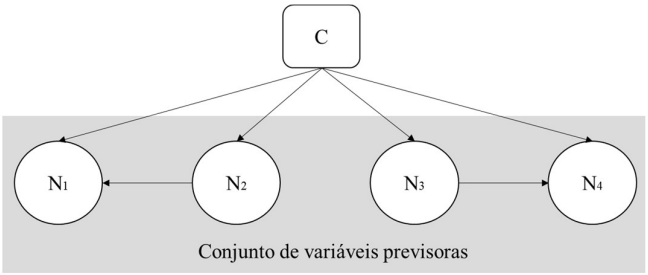

Figura 10 – Classificador TAN com 4 variáveis previsoras
Fonte: elaboração própria

A probabilidade a *posteriori* de $C=c_1$ para as variáveis previsoras $N = \{n_1, n_2, n_3, n_4\}$ do classificador TAN acima, é calculada pela equação 15.

$$P(c_1|n_1, n_2, n_3, n_4) = P(c_1).\frac{P(n_1|c_1, n_2).P(n_2|c_1).P(n_3|c_1).P(n_4|c_1, n_3)}{P(n_1, n_2, n_3, n_4, c_1) + P(n_1, n_2, n_3, n_4, c_2)}$$

(equação 15)

Um possível classificador Bayesiano TAN para o exemplo do vestibular está na Figura 11 e a probabilidade é calculada pela equação 16. Note que a variável de Nervosismo é também condicionalmente dependente de Concorrentes, o mesmo ocorre com Prova que é condicionalmente dependente de Desempenho.

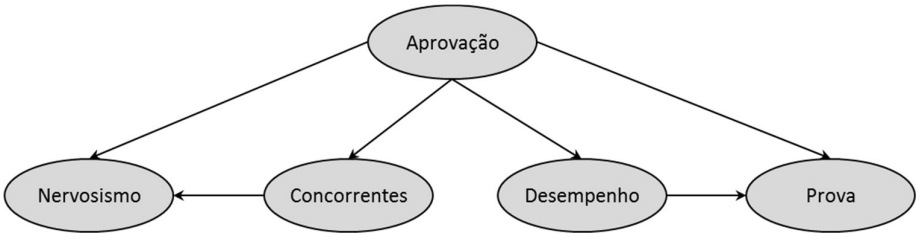

Figura 11 – Exemplo de classificador TAN para o exemplo vestibular
Fonte: elaboração própria

$$\mathrm{P}(\mathrm{Ap}=\mathrm{sim}|\mathrm{De}=\mathrm{alto},\mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta})=$$
$$=\frac{\mathrm{P}(\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{De}=\mathrm{alto}|\mathrm{Ap}=\mathrm{sim})\,.\mathrm{P}(\mathrm{Pr}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim},\mathrm{De}=\mathrm{alto}).\mathrm{P}(\mathrm{Ne}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim},\mathrm{Co}=\mathrm{alta}).\mathrm{P}(\mathrm{Co}=\ \mathrm{alta}|\mathrm{Ap}=\mathrm{sim})}{\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{sim})+\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{Não})}$$

(Equação 16)

2.1.2.4.3BN Augmented Naïve Bayes (BAN)

O classificador BN Augmented Naïve Bayes, Figura 12, é uma evolução do TAN. Nele a variável alvo continua sendo destacada das outras na determinação da estrutura da rede. De acordo com Cheng (1999), embora o critério de adicionar um novo arco seja o mesmo utilizando no classificador bayesiano TAN, a estrutura da rede não fica restrita a grafos em formato de árvore. Ao adicionar novos arcos à estrutura, as variáveis previsoras podem ter mais de dois nós pais e o classificador BAN pode ter uma complexidade maior do que o TAN, aumentando o tempo de processamento do modelo.

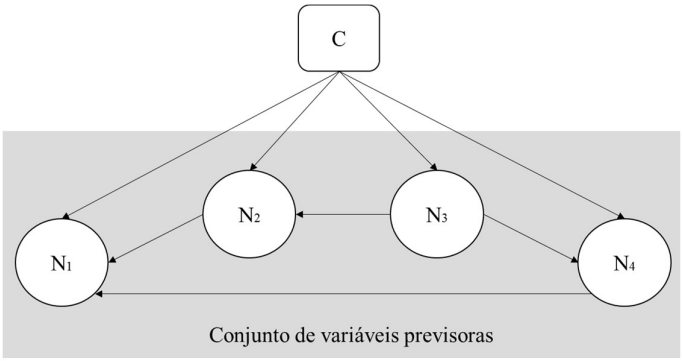

Figura 12 – Classificador BAN com 4 variáveis previsoras
Fonte: elaboração própria

A Figura 12 ilustra um classificador Bayesiano BAN com quatro variáveis previsoras. O conjunto de pais das variáveis previsoras são P_{N1} = {C, N_2, N_4}, P_{N2} = {C, N_1}, P_{N3} = {C} e P_{N4} = {C, N_3}. Os parâmetros das variáveis previsoras são obtidos por contagem de casos na amostra utilizada para determinar o classificador.

A probabilidade a *posteriori* da variável alvo, C=c_1, é dada pela equação 17.

$$P(c_1|n_1,n_2,n_3,n_4) = P(c_1).\frac{P(n_1|c_1,n_2,n_4).P(n_2|c_1,n_1).P(n_3|c_1).P(n_4|c_1,n_3)}{P(n_1,n_2,n_3,n_4,c_1)+P(n_1,n_2,n_3,n_4,c_2)}$$

(equação 17)

Um possível classificador BAN para o exemplo do vestibular está na Figura 13. Note que a variável Nervosismo é condicionalmente dependente de Concorrentes e Aprovação, e Prova é condicionalmente dependente de Nervosismo, Desempenho e Aprovação.

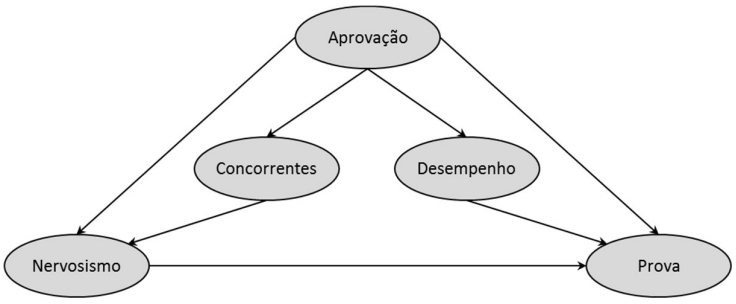

Figura 13 – Exemplo de classificador BAN para o exemplo vestibular
Fonte: elaboração própria

Nesse caso, a probabilidade de aprovação é dada pela equação 18.

$$\mathrm{P}(\mathrm{Ap}=\mathrm{sim}|\mathrm{De}=\mathrm{alto},\mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta})=$$
$$=\frac{\mathrm{P}(\mathrm{Ap}=\mathrm{sim}).\mathrm{P}(\mathrm{De}=\mathrm{alto}|\mathrm{Ap}=\mathrm{sim})\,.\mathrm{P}(\mathrm{Pr}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim},\mathrm{De}=\mathrm{alto},\mathrm{Ne}=\mathrm{médio}).\mathrm{P}(\mathrm{Ne}=\ \mathrm{médio}|\mathrm{Ap}=\mathrm{sim},\mathrm{Co}=\mathrm{alta}).\mathrm{P}(\mathrm{Co}=\ \mathrm{alta}|\mathrm{Ap}=\mathrm{sim})}{\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{sim})+\mathrm{P}(\mathrm{De}=\mathrm{alto},\ \mathrm{Pr}=\mathrm{médio},\mathrm{Ne}=\mathrm{médio},\mathrm{Co}=\mathrm{alta},\mathrm{Ap}=\mathrm{Não})}$$

(Equação 18)

2.1.3 Medidas estatísticas para avaliar a capacidade de classificadores

Independentemente da ferramenta estatística utilizada para desenvolver um classificador a partir de uma base de dados, que contém dados amostrais de uma população em estudo, é preciso avaliar a capacidade de o modelo discriminar os demais indivíduos da população.

De acordo com Sicsú (2010), há diferentes medidas estatísticas para isso como o KS - índice de Kolmogorov Smirnov, Coeficiente de Gini, D de Sommers, CAP (perfil de eficiência acumulado) e o AUROC (Area Under Receiver Operating Characteristic). Na área de análise de crédito o índice KS é um dos mais utilizados para avaliar o poder de discriminação dos modelos. No entanto, segundo Tomasella et al. (2008), quando a análise do modelo utiliza somente um desses índices, entre o AUROC e o KS, o primeiro apresenta um melhor resultado. Por isso, neste estudo será utilizado o AUROC.

2.1.3.1 AUROC – Area Under Receiver Operating Characteristic

Antes de apresentar o conceito de AUROC é preciso definir os conceitos de especificidade e sensitividade. Para isso utiliza-se uma matriz de classificação que pode ser definida como uma matriz nxn, onde n é o número de classes da variável alvo. As linhas dessa matriz possuem as classes dos valores reais e as colunas as classes dos valores previstos. Por exemplo, considere uma amostra de 100 alunos, classificados como bom desempenho e mau desempenho a partir do número de reprovações. Sabe-se que há 80 alunos com bom desempenho e 20 alunos com mau desempenho. Admita que um classificador indique a

existência de 70 alunos com bom desempenho e 30 alunos com mau desempenho. Para esse caso a matriz de classificação é mostrada na Tabela 4.

Tabela 4 – Matriz de classificação

		Previsto		
		Bom	Mau	Total
Realidade	Bom	65	15	80
	Mau	5	15	20
	Total	70	30	100

Fonte: elaboração própria

No exemplo, cinco alunos “Mau” foram classificados como “Bom” e quinze alunos “Bom” foram classificados como “Mau”. Houve coincidência em 65 alunos “Bom” e 15 alunos “Mau”. Nesse caso, há cinco casos falso positivos, ou seja, alunos classificados como “Bom”, mas que deveriam ser “Mau”.

Os conceitos de especificidade e sensibilidade podem ser ilustrados em uma matriz de classificação, conforme Tabela 5. A sensibilidade é medida pela proporção de casos “Bom” classificados corretamente. A especificidade é a proporção de casos “Mau” classificados corretamente. O conceito de 1 – especificidade é o falso positivo.

Tabela 5 – Especificidade e sensibilidade na matriz de classificação

		Previsto	
		Bom	Mau
Realidade	Bom	sensibilidade	1 - sensibilidade
	Mau	1-especificidade	especificidade

Fonte: elaboração própria

A medida estatística AUROC, ou apenas de ROC, é a área sob uma curva definida em um gráfico bidimensional cujo eixo da ordenada é a sensitividade do modelo e a abcissa é o valor complementar da especificidade, ou seja, 1 - especificidade. Na Figura 14, a curva 01 é resultado de um modelo sem discriminação e o AUROC é igual a 0,5. A curva 02 é o resultado de um modelo com alguma capacidade de discriminação e a curva 03 é o caso de um modelo com discriminação perfeita e o AUROC é igual a 1.

De acordo com Hosmer et al. (2013), a capacidade de discriminação do modelo depende do seu valor de ROC conforme indicado na Tabela 6.

Tabela 6– Capacidade de discriminação de um modelo de acordo com o valor de ROC

Valores do ROC	Capacidade de discriminação de um modelo
ROC = 0,5	Não discrimina
0,5 < ROC < 0,7	Baixa
0,7≤ ROC < 0,8	Satisfatório
0,8≤ ROC < 0,9	Excelente
ROC ≥ 0,9	Fora do comum

Fonte: Hosmer et al.,2013

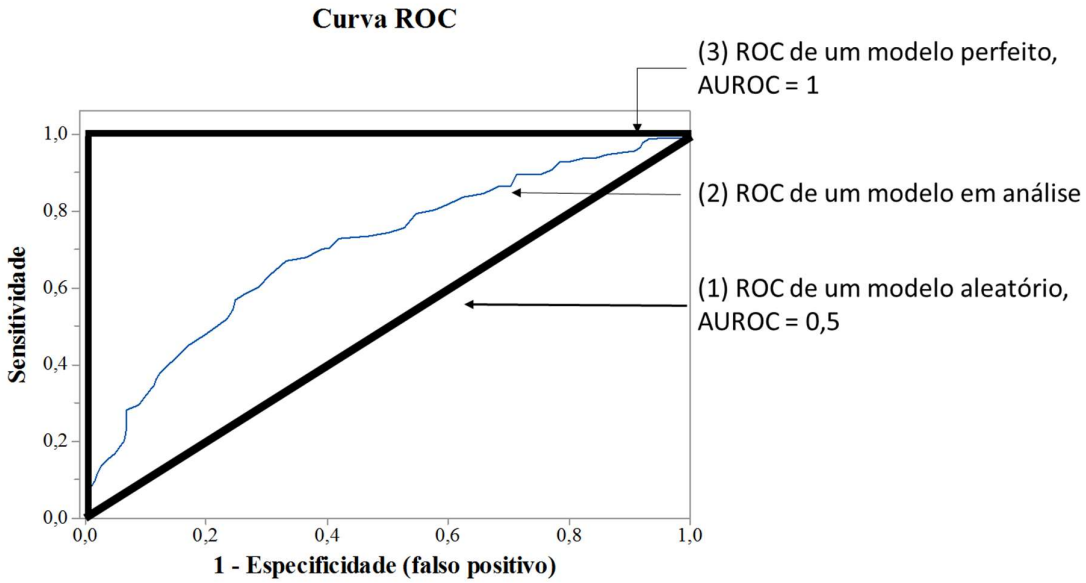

Figura 14 – Exemplo de curva ROC
Fonte: elaboração própria

2.2 Processo de KDD (Knowledge Discovery in Database)

Nas últimas décadas, observou-se um aumento significativo no volume de dados armazenados. Essa massa de dados disponível pode conter informações capazes de gerar novos conhecimentos para a humanidade. No entanto, é preciso de um método estruturado para fazer a busca. Neste estudo, será utilizado o conceito definido por Fayyad et al. (1996a) denominado KDD, *Knowledge Discovery in Database.* Ele é um processo não trivial para, em grandes

grupos de dados, identificar padrões válidos, novos, potencialmente úteis e possíveis de serem entendidos.

Na visão de Fayyad et al. (1996b) identificar padrões válidos significa ajustar um modelo aos dados coletados, encontrar uma estrutura dos dados, ou até mesmo descrever um grupo de dados. Fayyad et al. (1996c) descreve o processo recursivo de KDD, Figura 15, composto por 9 etapas para preparar os dados, buscar por padrões, avaliar o conhecimento e refiná-lo. A seguir serão definidas cada uma das etapas, exemplificando-as com o objeto de estudo desta dissertação.

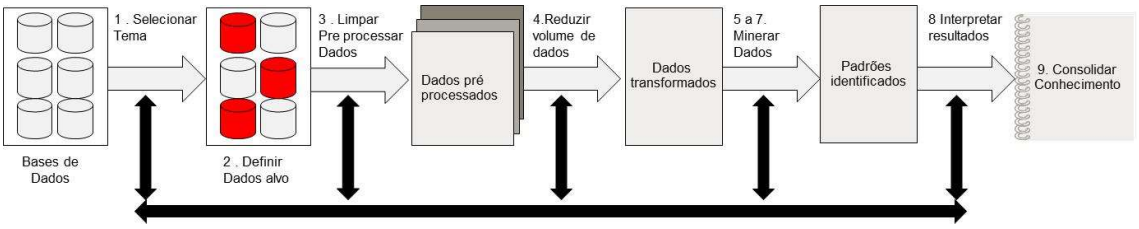

Figura 15 – Processo de KDD
Fonte: Adaptado de FAYYAD et al. 1996a

- **Etapa 01** – Selecionar tema. Essa etapa é o ponto de partida para o processo. O pesquisador que está aplicando-o precisa desenvolver um entendimento sobre o tema em análise e identificar quais são os objetivos do ponto de vista do usuário final. Neste estudo, o tema será o uso de classificadores para prever o desempenho de alunos da graduação da FGV/EAESP a partir de informações coletadas no processo de vestibular.

- **Etapa 02** – Definir dados alvo. Após definir o tema de estudo, é necessário selecionar quais informações serão analisadas. Neste estudo, serão utilizadas as bases dados com informações demográficas, formação educacional dos pais dos alunos, tipo de instituição de ensino médio (pública ou privada), notas do vestibular, maneira como o candidato teve conhecimento o vestibular da FGV/EAESP, número de reprovações por nota até o quarto semestre do curso.

- **Etapa 03** – Limpar e pré-processar os dados. Com o tema definido e as informações selecionadas, inicia-se o tratamento dos dados. É o momento para unificar as diferentes bases de dados, descartar informações incorretas ou suspeitas e buscar alternativas para tratar base de dados incompletas. Por exemplo, neste estudo, a base de dados do vestibular possui as informações de todos os candidatos e os dados daqueles não aprovados serão descartados.

- **Etapa 04** – Reduzir o volume de dados. Após limpar e pré-processar os dados, é preciso avaliar se é possível identificar padrões que permitirão reduzir o volume de dados e a complexidade. A base de dados final possui 114 informações que podem ser variáveis previsoras, volume excessivo para a construção de uma rede Bayesiana e um modelo de regressão logística, sabendo que o tamanho da amostra é de 1.370. Por exemplo, são eliminadas informações repetidas, como as notas do vestibular que estão em duas bases diferentes.
- **Etapa 05** – Alinhar os objetivos do processo de KDD (etapa 01) com um método de mineração de dados. Com a definição do tema a ser estudado, é necessário definir qual método de mineração de dados será utilizado. Neste estudo, a mineração de dados será por meio de classificadores Bayesianos e regressão logística.
- **Etapa 06** – Escolher o algoritmo de mineração de dados. Nessa etapa, é preciso definir quais algoritmos serão utilizados para identificar padrões nos dados. Para classificadores Bayesianos serão utilizados algoritmos para determinar o Naïve Bayes, TAN e BAN. Os softwares utilizados são o BayesiaLab, versão 5.4.3, para redes Bayesianas e o Minitab, versão 17, para regressão logística.
- **Etapa 07** – Realizar a mineração de dados. Com a definição do método e do algoritmo, a busca por padrões de interesse pode ser feita nos dados tratados na etapa 04. Neste estudo, espera-se identificar as variáveis previsoras que ajudam a prever se um aluno terá bom desempenho no curso de administração de empresas da FGV/EAESP.
- **Etapa 08** – Interpretar resultados da mineração. Esse passo consiste em interpretar os resultados obtidos da etapa anterior. Caso eles não estiverem de acordo com o esperado, é preciso voltar aos passos anteriores para identificar a razão do insucesso.
- **Etapa 09** – Consolidar o conhecimento descoberto. Nessa etapa, os resultados derivados da etapa 8 levam a uma conclusão, gerando conhecimento que poderá ser documentado para utilização em outras oportunidades.

3 Materiais e métodos

Nesta seção será apresentada uma descrição sucinta da metodologia utilizada para que o leitor tenha uma visão holística do processo.

3.1 Base de dados

Para este estudo, serão utilizadas base de dados disponibilizadas pela Secretaria de Ensino – Curso de Graduação e a pela Coordenadoria de Admissão aos Cursos Regulares da FGV-EAESP. As bases de dados são de candidatos e alunos do curso de administração de empresas ingressantes no período de 2009 a 2012. Elas estão divididas em dois grupos: informações de vestibular e informação da graduação.

- **Informações de vestibular -** Base de dados com o registro de todos os candidatos do vestibular da FGV-EAESP durante o período analisado. Essas bases possuem informações sócio demográficas dos candidatos, curso de graduação escolhido pelo candidato, desempenho nas provas do vestibular e respostas dos questionários preenchidos no momento da inscrição.
- **Informações de graduação** - A base de dados possui o registro dos alunos da graduação em administração de empresas no período analisado. Nesses registros estão a documentação dos alunos, número de matrícula, informações demográficas, notas finais das disciplinas cursadas e a indicação de quais disciplinas o aluno foi reprovado por nota.

3.2 Metodologia

A metodologia utiliza o processo de KDD como referência, sendo adaptado para o estudo. A Figura 16 ilustra a metodologia.

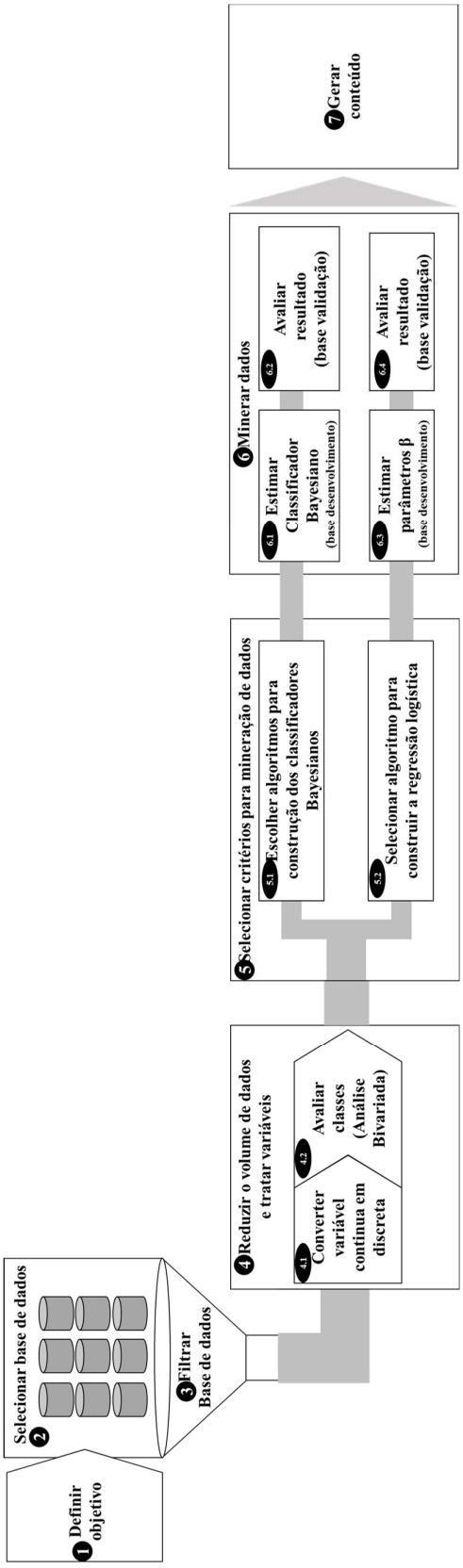

Figura 16 – Metodologia
Fonte: elaboração própria

Passo 01 – Definir objetivo. Este estudo utiliza classificadores Bayesianos e regressão logística para prever o desempenho de alunos da graduação da FGV/EAESP a partir de informações do vestibular.

Passo 02 – Selecionar base de dados. Essa etapa consiste em coletar as bases de dados com informações do vestibular e da graduação dos alunos no curso de administração de empresas da FGV-EAESP ingressantes entre 2009 e 2012. Deve-se garantir que há informações apenas dos alunos da graduação do curso de administração da FGV/EAESP que ingressaram através do vestibular. Todos os casos que não atendem a esses requisitos devem ser excluídos.

Passo 03 – Filtrar base de dados. Eliminar informações fora do objeto de estudo. É preciso avaliar se há casos que atendem aos requisitos do passo 02, mas que ainda não pertencem ao objeto de estudo, como é o caso de alunos que tiveram sua matrícula cancelada, por razões desconhecidas e não iniciaram o curso.

Passo 04 – Reduzir volume de dados e tratar variáveis. As bases de dados podem conter informações que não serão úteis para o estudo. Para elimina-las considerar quatro critérios:

1. **A informação não é discriminante:** informações como documentação e aquelas que são iguais para todos os casos da amostra, por exemplo, nome do curso se enquadra nesse critério.
2. **A informação é redundante:** informações que se repetem devem ser excluídas, mantendo apenas uma delas. Por exemplo, o município e estado onde se localiza a instituição de ensino do médio do candidato.
3. **A informação não está presente para todos os casos:** informações que não estão presentes para todos os casos e não podem ser estimadas por um algoritmo foram removidas. Por exemplo, a resposta à questão "Se foi aprovado no vestibular anteriormente, você chegou a fazer matrícula?" foi eliminada da base de dados porque ela está presente apenas nos questionários de 2009. Além disso, não há informações para inferir às respostas dos alunos a partir de 2010.
4. **As informações que podem gerar multicolinearidade:** informações que são combinação linear de outras devem ser removidas. Por exemplo, a média da primeira fase do vestibular é da média ponderada das provas da primeira fase, ou seja, é a combinação linear das provas de língua portuguesa, matemática, ciências humanas e língua inglesa. Nesse caso, as notas das provas devem ser mantidas e a média da primeira fase removida.

Após reduzir a base de dados, a discretização das variáveis contínuas é feita em duas etapas. A primeira etapa consiste em criar as classes das variáveis, no caso de notas de prova, serão inicialmente 10 classes, com variação de 1 ponto entre cada uma. A segunda etapa consiste em avaliar o número de elementos da amostra em cada classe criada. Se a classe possuir menos de 10% dos elementos da amostra e a variável for ordinal, a classe é fundida com a classe adjacente. Para variáveis nominais, as classes são fundidas de modo que a classe resultante tenha pelo menos 10% dos elementos da amostra.

Para toda as variáveis discretas, será feita uma análise bivariada utilizando o índice WOE (SICSÚ, 2010), calculado pela equação 19.

$$WOE = ln\frac{P(classe_i|classificaçãoA)}{P(classe_i|classificaçãoB)}$$

(Equação 19)

onde:

ln é o logaritmo natural da razão

$P(classe_i|classificaçãoA)$ é a quantidade de casos na amostra na classe i e na classificação A

$P(classe_i|classificaçãoB)$ é a quantidade de casos na amostra na categoria i e na classificação B

Após o cálculo do índice, serão unificadas categorias com base no índice de WOE. Para variáveis discretas nominais, as classes com WOE semelhantes são fundidas e para variáveis discretas ordinais, apenas classes adjacentes com WOE semelhantes são fundidas.

Após concluída a análise bivariada, as variáveis dummies são geradas. Elas são variáveis binárias que assumem os valores 0 e 1 e são utilizadas para representar classes de uma variável qualitativa. O número de variáveis dummies para uma variável com n classes é igual a n-1.

Passo 05 – Selecionar critérios para mineração de dados. O estudo utiliza dois softwares para o desenvolvimento dos classificadores. Para classificadores com redes Bayesianas será utilizado o BayesiaLab, versão 5.4.3, (CONRADY e JOUFFE, 2015) que possui os algoritmos para definição de classificadores Naïve Bayes, TAN e BAN.

Para regressão logística será utilizado o Minitab, versão 17. A pré-seleção das variáveis para o modelo de regressão logística será feita por três métodos Seleção *Forward*, Eliminação *Backward* e *Stepwise*.

- **Seleção *Forward*:** nesse método as variáveis são incluídas ao modelo uma de cada vez. A inclusão de uma nova variável ocorre até a capacidade de discriminação do modelo não sofrer melhoria (SICSÚ, 2010).
- **Eliminação *Backward*:** nesse método o modelo de regressão logística inicia com todas as variáveis disponíveis. Elas são eliminadas, uma a uma, até que a exclusão de uma variável comprometa a capacidade de discriminação do modelo (SICSÚ, 2010).
- ***Stepwise*:** nesse método as variáveis são adicionadas ao modelo gradativamente. Após a entrada de uma nova variável, o método verifica se as variáveis que foram selecionadas anteriormente podem ser eliminadas com a entrada da nova variável. O critério para inclusão ou não é o p-valor das variáveis (SICSÚ, 2010).

Para os classificadores Bayesianos a seleção das variáveis será feita em duas etapas. Baseando-se nos trabalhos de Karcher (2009) e Nogueira (2012), inicialmente será utilizado o conceito de entropia relativa e depois será feito um teste de independência de Chi^2 entre a cada potencial variável previsora e a variável alvo. Segundo Cover e Thomas (2006), no campo da teoria da informação, entropia é uma medida da incerteza de uma variável aleatória e ganho de informação é a redução da incerteza de uma variável quando se conhece o valor de outra variável. De acordo com Nogueira (2012), a entropia relativa, ou Divergência de Kullback-Leibler (DKL) varia de 0 a 1 e pode ser usada em redes Bayesianas para medir o ganho de informação que uma variável acarreta na outra. Neste estudo, será calculado a entropia relativa entre a variável alvo e cada potencial variável previsora, para identificar o ganho de informação da variável alvo devido a cada variável previsora. Aquelas que geram ganho de informação nulo serão eliminadas porque não adicionam informação à variável alvo; as outras serão mantidas inicialmente.

Pode ser que após essa primeira pré-seleção, ainda existam muitas variáveis previsoras e os trabalhos e livros pesquisados (KARCHER, 2009; NOGUEIRA, 2012; WITTEN e FRANK, 2005 e COVER E THOMAS, 2006) não indicam um valor mínimo de entropia relativa a ser usado como critério para eliminar variáveis com índice de DKL acima de zero. Assim, um teste de independência de Chi^2, para um p-valor de 15%, é feito entre a variável alvo e cada variável previsora com entropia relativa maior que zero a fim de concluir a seleção de variáveis.

Deste modo, para classificadores Bayesianos, serão eliminadas as variáveis previsoras que o teste de Chi^2 indicar independência entre elas e a variável alvo e as variáveis previsoras com entropia relativa, DKL, iguais ou muito próximo de zero.

- **Passo 06 – Minerar dados.** Nesta etapa, os classificadores são parametrizados a partir da amostra. Para isso, recomenda-se dividir a amostra em duas bases de dados numa proporção 70/30. A primeira, denominada base de desenvolvimento com 70% da amostra, é utilizada para estimar os parâmetros dos classificadores. A segunda, denominada base de validação com 30% da amostra, é utilizada para avaliar a qualidade do resultado do classificador. A variável alvo é a quantidade de reprovações por nota, sendo até uma reprovação o elemento é classificado como Bom e acima de uma reprovação como Mau. Todas as outras variáveis são potenciais variáveis previsoras.

A base de desenvolvimento será utilizada no Minitab para definir os parâmetros da regressão logística para os três métodos de seleção de variável, considerando um p-valor de 15%.

Em seguida, é realizado o teste de Hosmer-Lemeshow para testar a aderência dos modelos à amostra e os indicadores ROC, TAT, taxa de acerto total, TAB, taxa de acerto de Bom e o TAM, taxa de acerto de Mau, serão utilizados para avaliar a capacidade de discriminação dos modelos. O cálculo do TAT, TAB e o TAM é ilustrado considerando os dados da Tabela 4. O TAT é o percentual de casos classificados corretamente na amostra analisada, nesse caso, o TAT = (65+15/100) = 80%, o TAB é o percentual de casos classificados corretamente como Bom, ou seja, TAB = 65/ 80 = 81,25% e o TAM é o percentual de casos classificados corretamente como Mau, ou seja, TAM = 15/20 = 75%.

Os modelos estimados serão testados na base de validação usando os índices AUROC, TAB, TAM e TAT. O modelo de regressão logística com melhor indicador e menor complexidade será selecionado. A complexidade do modelo de regressão logística é proporcional ao número de variáveis previsoras.

A base de desenvolvimento também será utilizada estimar os classificadores Bayesianos Naïve Bayes, TAN e BAN. Como o modelo de regressão logística, os indicadores ROC, TAT, TAB e o TAM serão utilizados para avaliar a poder de discriminação desses classificadores. Aquele com melhor poder de discriminação será o escolhido. Após identificar o melhor classificador Bayesiano, as variáveis previsoras selecionadas no modelo de regressão logística serão utilizadas para estimar esse mesmo classificador Bayesiano para avaliar se há convergência no resultado dos classificadores.

A base validação será utilizada nos classificadores Bayesianos, como foi feito no modelo de regressão logística.

- **Passo 07 – Geração de conteúdo.** Após a validação dos algoritmos classificadores, será possível identificar e concluir quais variáveis previsoras influenciam no desempenho do aluno.

4 Apresentação dos resultados / tratamento das informações

A apresentação dos resultados iniciará a partir do passo 02 da metodologia porque a definição do objetivo do estudo já foi feita na seção 01.

Passo 02 - Selecionar base de dados

As duas bases mencionadas na 3.1 foram constituídas a partir de diferentes bases de dados disponibilizadas pelos departamentos responsáveis. A Coordenadoria de Admissão aos Cursos Regulares forneceu três bases de dados para construir a base de informações de vestibular: dados de inscrição, respostas do questionário e notas do vestibular.

a) Dados de inscrição: a base contém as seguintes informações do candidato: demográficas, documentação, endereço e dados sobre o ensino médio. Além disso, há informações de candidatos interessados no curso de Administração Pública, porque até o vestibular do primeiro semestre de 2011, era possível escolher entre os cursos Administração Pública e Administração de Empresas. O total de inscritos no período é de 25.904 candidatos dos quais 1.370 se matricularam.
b) Repostas do questionário: no momento da inscrição o candidato deve preencher um questionário, que vem se alterando ao longo dos anos. No entanto, para o período analisado há 7 questões de 8, que foram mantidas e são listadas a seguir. O número questionários preenchidos foi de 25.900.
    1) Você já prestou vestibular para a FGV-EAESP anteriormente? Respostas possíveis: Sim/Não
    2) Para que instituições você já prestou ou pretende prestar vestibular neste ano? Reposta livre para o candidato indicar o nome das instituições
    4) Com relação a estudar na FGV-EAESP, você se considera: Repostas possíveis: muito indeciso, indeciso, decidido, muito decidido
    5) Em qual das faixas abaixo, você estima estar a soma da renda mensal total da sua família? Respostas possíveis: Abaixo de R$4.650, De R$4.651 a R$9.300, De R$ 9.301 a R$18.600, De R$18.601 a R$27.900, De R$27.901 a R$37.200 e Acima de R$37.200
    6) Qual é o nível de instrução do seu pai? Repostas possíveis: Não frequentou a escola, Ensino fundamental completo, Ensino médio completo, Graduação completa, Pós-

graduação completa (especialização, MBA), Pós graduação completa (mestrado) e Pós-graduação completa (doutorado, PhD).

7) Qual é o nível de instrução de sua mãe? Repostas possíveis: Não frequentou a escola, Ensino fundamental completo, Ensino médio completo, Graduação completa, Pós-graduação completa (especialização, MBA), Pós graduação completa (mestrado) e Pós-graduação completa (doutorado, PhD).
8) Você observou divulgação sobre o vestibular da FGV-EAESP em algum dos seguintes meios? Repostas possíveis: Mala Direta, Internet, Jornal, TV aberta, TV fechada, Cinema, Panfleto, Cartaz, Evento "Conversa com Administradores", Feiras, Palestras no Colégio, Não vi nenhuma divulgação e Outros.

c) Notas do vestibular: a principal informação desta base de dados é o desempenho do candidato nas provas do vestibular da FGV/EAESP. O total de candidatos é de 18.849, menor do que a quantidade da base anterior, porque acredita-se que alguns candidatos inscritos não realizaram as provas.

A Secretaria de Ensino – Curso de Graduação forneceu quatro bases de dados para a base de informações da graduação: dados do vestibular, dados dos alunos da graduação, dados do perfil do aluno de graduação e notas das disciplinas da graduação.

a) Dados do vestibular: a base de dados contém informações do desempenho no vestibular dos alunos matriculados no curso de administração de empresas da FGV-EAESP. Nessa base há 2.651 registros, incluindo alunos que realizaram a matrícula, porém cancelaram-na por razões desconhecidas.
b) Dados dos alunos da graduação: a base de dados possui os dados do aluno na graduação, a sua situação atual, podendo ser ativo se ainda está cursando, concluído se já é formado, evadido e jubilado, se foi eliminado do curso a partir de normas da FGV/EAESP e cancelado. De acordo com o Secretaria de Graduação, alunos podem ser classificados como cancelados nas seguintes situações: abandono de curso, desistência do curso, cancelamento de matrícula, novos ingressantes que não efetuaram matrícula, embora tenham sido aprovados. Nessa base há 2.013 registros. É uma quantidade menor do que o número de casos na base dados do vestibular, porque alunos com matrícula cancelada foram excluídos. No entanto, isso não afeta a análise do estudo, porque os alunos cancelados não cursaram o período completo para análise.
c) Dados do perfil do aluno de graduação: a base de dados possui informações demográficas do aluno, documentação e informações sobre o ensino médio. Há uma sobreposição com a base dado de inscrição, porém é uma redundância para garantir que não haverá registros

incompletos para a amostra em estudo. Nessa base há 2.673 registros, porque há casos repetidos, que foram eliminados, e dados sobre os alunos que ingressaram em Administração Pública nos vestibulares do ano de 2009.

d) Notas das disciplinas da graduação: esta base com 2.013 registros possui as notas de todas as disciplinas cursadas pelo aluno durante a graduação. Há indicação se o aluno foi reprovado por nota ou não.

As sete bases de dados podem ser resumidas na Tabela 7 e o número total de informações disponíveis é de 114, sem considerar repetições, como pode ser visto no apêndice A. Elas foram organizadas para iniciar o processo de aglutinação a ser detalhado no passo 03.

Tabela 7 – Número de informações e registros nas bases de dados disponibilizadas para o trabalho

Fonte	Nome da base de dados	Número de informações	Número de registros
Coordenadoria de Admissão aos Cursos Regulares	a)Dados de inscrição	49	25.904
	b)Reposta do questionário	15	25.900
	c)Notas do vestibular	35	18.849
Secretaria de Ensino – Curso de Graduação	a)Dados do vestibular	24	2.651
	b)Dados do aluno da graduação	16	2.013
	c)Dados do perfil do aluno de graduação	33	2.673
	d)Notas das disciplinas da graduação	15	2.013
Total de informações com repetição		187	

Fonte: elaboração própria

Passo 03 – Filtrar base de dados

Antes de filtrar a base de dados, foi identificada uma informação comum entre as sete bases para uni-las. Embora, acreditasse que o elo de ligação entre as bases de dados do vestibular e as bases da graduação fosse ou o RG ou o CPF, a melhor informação para isso foi o nome da pessoa. Nas bases de dados, havia alguns casos de um candidato ter número de CPF ou RG diferentes. Por exemplo em um caso hipotético, na base do vestibular há apenas um candidato chamado Alex Kuribara com um número de CPF e RG. Esse candidato foi aprovado e na base de dados da graduação o CPF ou RG não são iguais aos números encontrados na base de vestibular. Pode-se pensar em candidatos homônimos, mas isso foi descartado porque nos casos com esse tipo de inconsistência, não foram encontrados dois registros com mesmo nome nas bases.

Após a identificação do elo de ligação, foi criada uma base intermediária de 2013 registros, utilizando como referência os alunos existentes na base Notas das disciplinas da graduação. Essa base foi a referência porque é preciso ter a informação de alunos que ingressaram em administração de empresas na FGV/EAESP no período analisado e concluíram ou estão concluindo o curso. Como resultado, foi constituída uma base intermediária com 2.013 casos composta por dados do vestibular e da graduação.

Após a sua constituição, foram removidos registros de alunos com matrícula cancelada e casos de alunos que não ingressaram por meio do processo do vestibular. O resultado foi uma base de dados final com 1.370 alunos e 114 potenciais variáveis previsoras. Esses alunos cursaram Administração de Empresas da FGV/EAESP até pelo menos o semestre 04 e ingressaram por meio do processo de vestibular e a quantidade é coerente com número teórico de vagas conforme detalhado na Tabela 8.

Tabela 8 – Comparativo entre o número de registro da base de dados e o número de vagas teóricas disponíveis para o curso de administração

Períodos	Registros na base de dados	Número de vagas teórico
2009	**308**	**300**
Semestre 01	152	150
Semestre 02	156	150
2010	**309**	**300**
Semestre 01	149	150
Semestre 02	160	150
2011	**358**	**350**
Semestre 01	151	150
Semestre 02	207	200
2012	**395**	**400**
Semestre 01	193	200
Semestre 02	202	200
Total	**1.370**	**1.350**

Fonte: elaboração própria

A quantidade de alunos dessa base é consistente com o número teórico de vagas oferecidas nos oito semestres entre 2009 a 2012, porque nos 5 primeiros semestres analisados eram oferecidas 150 vagas e nos últimos 3 semestres, 200 vagas, totalizando 1.350. De acordo com a FGV/EAESP, pode haver períodos que são matriculados mais alunos do que a quantidade teórica de vagas, como pode ser comprovado com a existência de 20 alunos além do número teórico de vagas disponíveis. No passo 04 será feita redução no volume da base de dados, avaliando as potenciais variáveis previsoras.

Passo 04 – Reduzir volume de dados e tratar variáveis

Os quatro critérios foram aplicados para reduzir o volume de dados.

1. **A informação não é discriminante:** todos os alunos analisados são solteiros, foram aprovados na primeira e segunda fase do vestibular, pagaram a taxa de inscrição do vestibular, ingressaram na FGV/EAESP pelo processo de vestibular, não eram treineiros no ano de ingresso, se matricularam no curso de administração de empresas, possuem número de telefone, correio eletrônico e nenhum é ex-aluno da instituição. Como essas informações não diferenciam os alunos, elas foram eliminadas. Além disso, dados como sexo, RG, CPF, Número de inscrição no vestibular, código do aluno, turma de referência, semestre e ano de ingresso entre outras foram eliminadas porque são para caracterização do aluno. A lista completa de informações com sua respectiva classificação está no apêndice A.
2. **A informação é redundante:** Um exemplo de informações repetidas são o endereço, estado, cidade e CEP porque elas indiretamente indicam a localização geográfica do aluno. Neste estudo, o interesse é saber a região do Brasil de onde o aluno é proveniente, como o primeiro dígito do CEP fornece essa informação, ele foi utilizado e as outras informações referentes à localização geográfica foram eliminadas.
3. **A informação não está presente para todos os casos:** Quando a variável não contém informação para todos os casos, ela foi eliminada, por exemplo nome do curso pré-vestibular onde o candidato estudou. Nesse caso, ou havia o nome ou o campo estava em branco, não indicando se o candidato não cursou, ou não preencheu. O campo foi eliminado porque não há um método para inferir se o aluno realizou ou não o curso sem que o candidato fosse consultado ou fossem verificadas todas as bases de registro dos cursos pré-vestibular do país. Outro caso de informação removida são as respostas de questões pertencentes apenas ao questionário dos vestibulares de 2009.
4. **As Informações que podem gerar multicolinearidade:** as médias das provas da primeira e segunda fase foram removidas e foram mantidas as notas das provas porque as médias são combinação lineares das notas das provas aplicadas em cada fase.

Após aplicar esses critérios na base de dados, chegou-se a dezesseis potenciais variáveis previsoras, como mostra a Tabela 9. Elas estão detalhadas no apêndice B e serão analisadas na próxima etapa.

Tabela 9 - Lista de potenciais variáveis previsoras

Potenciais variáveis previsoras	Qualitativa / Quantitativa	Tipo
Prova 1bruta – Matemática Fase 01	Quantitativa	Contínua
Prova 2bruta – Português Fase 01	Quantitativa	Contínua
Prova 3bruta – Inglês Fase 01	Quantitativa	Contínua
Prova 4bruta – Ciências Humanas Fase 01	Quantitativa	Contínua
Prova 5bruta – Matemática Aplicada Fase 02	Quantitativa	Contínua
Prova 6bruta – Redação Fase 02	Quantitativa	Contínua
Idade	Quantitativa	Contínua
Região	Qualitativa	Nominal
Tipo de instituição ensino médio	Qualitativa	Nominal
Questão 01	Qualitativa	Nominal
Questão 04	Qualitativa	Ordinal
Questão 05	Qualitativa	Ordinal
Questão 06	Qualitativa	Ordinal
Questão 07	Qualitativa	Ordinal
Questão 08	Qualitativa	Nominal
Período no ensino médio	Qualitativa	Nominal

Fonte: elaboração própria

A data de nascimento foi convertida para idade do aluno quando ingressou no curso. O CEP foi convertido para a variável região, utilizando o seu primeiro dígito para identificar a região de onde cada aluno é proveniente.

As variáveis Provas inicialmente foram discretizadas utilizando classes com variação de 1 ponto entre elas. Por exemplo, a Prova 6 foi discretizada inicialmente em 8 classes. Ao avaliar a distribuição da amostra nas categorias, nota-se que 4 delas possuem menos de 10% do total da amostra. Quando isso ocorre, Sicsú (2010) recomenda que as categorias sejam fundidas para conterem pelo menos 10% da amostra. Aplicando esse critério, as classes Entre 2 e 3 inclusive e Entre 3 e 4 inclusive foram fundidas em uma nova classe chamada Entre 2 e 4 inclusive. O mesmo método foi feito para as outras classes e a variável Prova 6 bruta – Redação Fase 02 ficou com 5 classes, conforme Tabela 10.

Tabela 10 – Discretização da variável Prova 6 bruta – Redação Fase 02

Classes da variável Prova 6 bruta – Redação Fase 02	DP ≤ 1	DP >1	Classes Agrupadas
Entre 2 e 3 inclusive	1%	2%	Entre 2 e 4 inclusive
Entre 3 e 4 inclusive	6%	7%	
Entre 4 e 5 inclusive	15%	19%	Entre 4 e 5 inclusive
Entre 5 e 6 inclusive	28%	29%	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	32%	28%	Entre 6 e 7 inclusive
Entre 7 e 8 inclusive	15%	14%	Acima de 7
Entre 8 e 9 inclusive	3%	1%	
Acima de 9	0%	0%	
Total Geral	**100%**	**100%**	

Fonte: elaboração própria

Para as variáveis qualitativas nominais houve o agrupamento de classes considerando que cada classe deve ter pelo menos 10% da amostra. Por exemplo, aplicando o critério a variável Região passou a ter 3 classes, conforme Tabela 11.

Tabela 11 – Fusão das classes da variável Região

Classes da variável Região	DP ≤ 1	DP >1	Classes Agrupadas
Grande SP	70%	74%	Grande SP
Outras regiões	9%	9%	Outros
Outro País	3%	2%	
SP Interior	19%	15%	SP Interior
Total Geral	**100%**	**100%**	

Fonte: elaboração própria

Para variáveis qualitativas ordinais houve uma fusão considerando a mesma distribuição mínima por classe, porém respeitou-se a ordem das classes para fundi-las. Após aplicar esse critério, a variável Questão 06 ficou com 5 classes como pode ser vista na Tabela 12.

Tabela 12 – Fusão das classes da variável Questão 06

Classes da variável – Questão 06/nível de instrução pai	**DP ≤ 1**	**DP >1**	**Classes Agrupadas**
Não freqüentou a escola	0%	0%	NF_Grau1
Ensino fundamental completo (1o grau)	2%	2%	
Ensino médio completo (2o grau)	12%	14%	Ensino médio completo (2o grau)
Graduação completa (ensino superior)	48%	49%	Graduação completa (ensino superior)
Pós-graduação completa (especialização)	25%	22%	Pós-graduação completa (especialização)
Pós-graduação completa (mestrado)	7%	8%	Mestrado_Doutorado
Pós-graduação completa (doutorado, PhD)	6%	6%	
Total Geral	**100%**	**100%**	

Fonte: elaboração própria

As classes Não frequentou a escola e Ensino fundamental completo (1º grau) foram fundidas formando a classe NF_Grau1. O mesmo ocorreu com as classes Pós-graduação completa (mestrado) e Pós-graduação completa (doutorado, PhD) que foram fundidas na classe Mestrado_Doutorado. A análise de classe de todas as variáveis está no apêndice B.

Após discretizar as variáveis contínuas e realizar uma primeira fusão das classes das variáveis, utilizou-se o índice de WOE para avaliar a capacidade de discriminação das classes. Classes de variáveis com WOE semelhantes foram fundidas respeitando os critérios para variáveis ordinais e nominais. A Tabela 13 apresenta o resultado da análise do índice WOE para a variável Prova 6 bruta - Redação.

As classes Abaixo de 4 inclusive e Entre 4 e 5 inclusive da variável Prova 06 foram fundidas porque possuem WOE semelhantes. Com isso, elas formaram uma nova classe denominada Abaixo de 5 inclusive. Aplicando raciocínio análogo para as outras classes, a variável Prova 6 passou a ter três classes.

Tabela 13 – Análise bivariada para a variável Prova 6 bruta - Redação

	Classificação				
Prova 6 bruta - Redação	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 4 inclusive	7%	9%	0,78	-0,254	Abaixo de 5 inclusive
Entre 4 e 5 inclusive	15%	19%	0,81	-0,214	
Entre 5 e 6 inclusive	28%	29%	0,96	-0,043	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	32%	28%	1,15	0,135	Acima de 6
Acima de 7	18%	15%	1,19	0,175	
Total	100%	100%			

Fonte: elaboração própria

A mesma análise foi aplicada para a variável Região, conforme Tabela 14. No caso da variável Região, os índices WOE das classes não são semelhantes por isso as classes são mantidas.

Tabela 14 - Análise bivariada para a variável Região

	Classificação				
Região	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Grande SP	70%	74%	0,94	-0,062	Grande SP
SP Interior	19%	15%	1,24	0,212	SP Interior
Outros	12%	11%	1,09	0,084	Outros
Total	100%	100%			

Fonte: elaboração própria

Para a variável Questão 06, a análise do WOE permitiu fundir as classes NF_Grau1 e Ensino médio completo (2º grau) formando a classe NF_Grau1_Grau2. Com isso, ela passou a ter 4 classes, conforme Tabela 15.

Tabela 15 – Análise bivariada para a variável Questão 06

	Classificação				
Questão 06	**DP≤1**	**DP > 1**	**Bivariada**	**WOE**	**Novas Classes**
NF_Grau1	2%	2%	0,87	-0,138	NF_Grau1_Grau2
Ensino médio completo (2o grau)	12%	14%	0,88	-0,124	
Graduação completa (ensino superior)	48%	49%	0,98	-0,019	Graduação completa (ensino superior)
Pós-graduação completa (especialização, MBA)	25%	22%	1,17	0,155	Pós-graduação completa (especialização, MBA)
Mestrado_Doutorado	13%	14%	0,94	-0,058	Mestrado_Doutorado
Total	100%	100%			

Fonte: elaboração própria

A análise do índice WOE das demais variáveis não apresentadas estão no apêndice C. Após verificar a capacidade de discriminação das classes das variáveis, a próximo passo é criar as variáveis dummies. Por exemplo, para a variável Prova 06 foram criadas 2 variáveis dummies, a VP6-D5a6, para a classe Entre 5 e 6 inclusive, e a VP6-DA6, para a classe Acima de 6. No total foram criadas 47 variáveis dummies a serem utilizadas nos passos seguintes.

Passo 05 – Selecionar critérios para mineração de dados

O Minitab, versão 17, é o software utilizado para o desenvolvimento do modelo de regressão logística. Serão feitos três modelos a partir dos três critérios de seleção de variável: Seleção *Forward*, Eliminação *Backward* e *Stepwise*, utilizando um p-valor de 15%. O teste de aderência do modelo será feito pelo teste Hosmer-Lemershow e a capacidade de discriminação será avaliada utilizando os índices ROC, TAB, TAM e TAT.

O BayesiaLab, versão 5.4.3 (CONRADY et JOUFFE, 2015) será utilizado para desenvolver os classificadores Bayesianos. Ele possui algoritmos que determinam as estruturas de rede dos classificadores Naïve Bayes, TAN e BAN e os parâmetros das variáveis a partir da base de dados. Além disso, calcula a entropia relativa entre as variáveis previsoras e a variável alvo e realiza o teste de independência de Chi^2. Assim, ele será utilizado para determinar os três classificadores Bayesianos e um quarto classificador será estimado usando as variáveis selecionadas no modelo de regressão logística para avaliar se a diferença de critérios para seleção de variáveis impacta a capacidade de discriminação do modelo. Essa capacidade será avaliada utilizando os índices ROC, TAB, TAM e TAT.

Passo 06 – Minerar dados

Os modelos foram desenvolvidos em duas etapas. A primeira consistiu em determiná-los utilizando uma base de desenvolvimento, composta por 70% da base de amostras, ou seja, com 959 registros. A segunda consistiu em testar o modelo em uma base de validação, composta por 30% da base de amostras, ou seja, 411 registros. Em todos os casos, a variável alvo foi a Bom aluno que pode assumir dois valores Bom e Mau. Ela indica se o aluno teve mais de uma reprovação por nota durante os quatro semestres do curso. Se ele tiver mais de uma DP, o valor da variável é Mau, caso contrário, o valor é igual a Bom. A divisão da base em desenvolvimento e validação foi arbitrárias e 50,1% dos casos da base de desenvolvimento tem bom desempenho e 55,1% dos casos da base de validação tem bom desempenho.

Modelo de regressão logística

O modelo de regressão logística foi desenvolvido a partir das 47 variáveis dummies definidas no passo 04. O primeiro modelo considerou todas as variáveis e o resultado do teste de significância indicou 35 variáveis com p-valor acima de 15%. Isso indica que elas rejeitam H_0: $\beta_n = 0$, n=1,..,35, ou seja, considerando todas as variáveis no modelo, algumas delas não contribuem para estimar a probabilidade de um aluno ter bom desempenho.

Após esse resultado, foram usados três critérios para seleção de variáveis Seleção Forward, Eliminação Backward e Stepwise, para um p-valor de 15%; obtendo três modelos de regressão logística respectivamente. A aderência desses modelos à base de desenvolvimento foi analisada a partir do teste de Hosmer-Lemershow. Como a Tabela 16 indica, os três modelos tiveram um p-valor acima de 15%, indicando que os modelos são aderentes à base de desenvolvimento.

Tabela 16 – p-valor para o teste de Hosmer-Lemershow

Critérios seleção de variáveis	P-valor (%) para Hosmer-Lemershow
Stepwise	84%
Seleção Forward	84%
Eliminação Backward	74%

Fonte: elaboração própria

O próximo passo no desenvolvimento de um modelo de regressão logística é avaliar sua capacidade de discriminação. O índice de ROC foi calculado para os três modelos considerando a base de desenvolvimento e a base de validação. Além disso, os índices de taxa de acerto total, TAT, taxa de acerto de Bom, TAB, e taxa de acerto de Mau, TAM, foram calculados. Os resultados estão resumidos na Tabela 17 e com base na Tabela 6, Seção 2.1.3.1 os modelos possuem uma baixa capacidade de discriminação porque o índice ROC é menor que 70% para todos os casos. O índice TAB corrobora para isso, porque na base de desenvolvimento ele está entre 66,9% e 68,5% e na base de teste entre 64,1% e 67,5%. Essa piora do índice na base de teste, sinaliza que os modelos tiveram *overfitting* à base de desenvolvimento. O TAM apresenta valores próximo de 60% e o TAT também apresentam valores próximos a 65%, indicando que o modelo classificaria incorretamente 35% dos casos analisados.

Tabela 17 – Índices para avaliar a capacidade de discriminação dos modelos de regressão logística

	Base Desenvolvimento			Base Validação		
Índice	Stepwise	Seleção Forward	Eliminação Backward	Stepwise	Seleção Forward	Eliminação Backward
ROC	69,0%	69,0%	69,0%	69,0%	69,0%	69,0%
TAB[1]	68,5%	68,5%	66,9%	64,1%	64,1%	67,5%
TAM[1]	58,7%	58,7%	60,0%	63,4%	63,4%	62,4%
TAT[1]	63,8%	63,8%	63,6%	63,7%	63,7%	65,0%

1) TAB é taxa acerto BOM, TAB é taxa acerto MAU e TAT é taxa acerto Total

Fonte: autor

Embora o critério sugerido por Hosmer e Lemershow (2013) indique que os modelos apresentam baixa capacidade de discriminação, eles mostram quais variáveis previsoras podem indicar se o aluno de graduação terá um bom desempenho. Como todo os modelos obtiveram índices semelhantes, foi escolhido modelo baseado no critério Stepwise porque apresentou resultado idêntico ao modelo baseado na Seleção Forward e possui menos variáveis previsoras que o modelo baseado na Eliminação Backward, conforme
Tabela 18. A escolha por um modelo com menos variáveis previsoras busca reduzir a sua complexidade.

Tabela 18 – Quantidade de variáveis dummies previsores por modelo de regressão logística

Modelo de Regressão Logística	**Quantidade de variáveis dummies previsoras**
Reg. Log. - Stepwise	14
Reg. Log. - Seleção Forward	14
Reg. Log. - Eliminação Backward	15

Fonte: elaboração própria

Tabela 19 apresenta a lista de variáveis previsoras e as respectivas variáveis dummies do modelo de regressão logística selecionado. Embora a sua capacidade de discriminação seja baixa, os coeficientes de regressão logística e a razão de chance das variáveis indicam quais variáveis contribuem positivamente e negativamente para identificar se um aluno que terá bom desempenho na graduação.

Sem utilizar o conhecimento de especialistas, os sinais dos coeficientes do modelo de regressão parecem coerente ao pensar que alunos com boas notas no vestibular tendem a ter melhor desempenho, assim como imaginar que pessoas dispostas a sair de suas cidades para estudar em uma faculdade na cidade de São Paulo estão interessadas pelo curso refletindo no

seu desempenho acadêmico. A variável relacionada com a idade do aluno no ano de ingresso parece ter coerência em seu sinal porque indica que quanto maior a idade no ano de ingresso, menor a probabilidade de ter bom desempenho. Intuitivamente, parece fazer sentido porque pessoas que ingressam no ensino superior com 17 ou 18 anos geralmente tiveram um alto desempenho acadêmico no ensino médio e tendem a manter isso na faculdade. Essa variável é coerente com a variável de número de vezes que o aluno prestou vestibular porque alunos que não estão preparados geralmente prestam o vestibular mais de uma vez até serem aprovados, refletindo na idade do ano de ingresso. O coeficiente do tipo de instituição de ensino, parece incoerente porque costuma-se dizer que alunos de escola privada são melhores do que alunos de escolas públicas. No entanto, de acordo com o coordenador da graduação de cursos da FGV/EAESP, alunos de escola pública tendem a ter bolsa de estudos e a condição para manter a bolsa é ter bom desempenho acadêmicos, sendo uma das razões para explicar o sinal do coeficiente.

A razão de chances completa a análise do modelo de regressão logística. Iniciando pelos resultados das provas do vestibular, os resultados mostram que um aluno com nota na prova de matemática na fase 01 entre 7 e 9, nota na prova de inglês entre 5 e 6, nota na prova de ciências humanas entre 7 e 8, nota na prova de matemática aplicada da fase 02 entre 5 e 6 ou acima de 8 e nota na redação acima de 6 tem maior chance de ter um bom desempenho no curso de administração de empresas da FGV/EAESP. O resultado indica que um conhecimento na língua inglesa que resulte em uma nota entre 5 e 6 aumenta em duas vezes as chances de o aluno ter um bom desempenho.

Alunos provenientes do interior de São Paulo tem uma maior chance de ter um bom desempenho no curso, entretanto, quanto mais velho no ano de ingresso, menor suas chances de ter um bom desempenho, como pode ser visto na razão de chances das variáveis VIE-D18a19 e VIE-D19a20, 0,74 e 0,55 respectivamente. Essas razões de chance indicam que um aluno com idade no ano de ingresso entre 18 e 19 tem 26% chance a menos de ter um bom desempenho se comparado com alunos que tinha entre 17 e 18 anos. Esse percentual sobe para 44% se o aluno tinha entre 19 e 20 anos.

Alunos provenientes de instituições de ensino médio particular e com renda entre R$18,6 mil e R$29,7mil também tem suas chances de ter um bom desempenho diminuídas. Por fim, alunos que conheceram o vestibular por meio da internet possuem uma maior chance de ter um bom desempenho e alunos que prestaram o vestibular mais de uma vez têm suas chances de ter um bom desempenho reduzidas.

É importante lembrar que o grau de instrução dos pais do aluno e o seu grau de decisão no momento de prestar o vestibular, capturado por meio da questão 04 do questionário de inscrição do vestibular não influenciam no seu desempenho futuro.

Tabela 19 – Lista de variáveis previsoras do modelo de regressão logística

Grupo de Variável Previsora	Variáveis previsoras e constante do modelo de regressão logística	Classe da variável previsora	Código da variável dummy	Coeficiente Regressão Logística	Razão de chances
	Constante do modelo de regresão logística	Não aplicável	Não aplicável	-0,567	
Provas do vestibular	Matemática - Fase 01	Nota entre 7 e 8 inclusive	VP1-D7a8	0,373	1,45
		Nota entre 8 e 9 inclusive	VP1-D8a9	0,294	1,34
	Inglês - Fase 01	Nota entre 5 e 6 inclusive	VP3-D5a6	0,711	2,04
	Ciências Humanas - Fase 01	Nota entre 7 e 8 inclusive	VP4-D7a8	0,417	1,52
	Matemática Aplicada - Fase 02	Nota entre 5 e 6 inclusive	VP5-D5a6	0,554	1,74
		Nota acima de 8	VP5-DA8	0,395	1,48
	Redação - Fase 02	Nota acima de 6	VP6-DA6	0,443	1,56
Informações socio-econômicas	Região onde morava durante ensino médio	São Paulo Interior	VRE - DSPI	0,336	1,40
	Idade do aluno no ano de ingresso	Entre 18 e 19 inclusive	VIE - D18a19	-0,300	0,74
		Entre 19 e 20 inclusive	VIE - D19a20	-0,604	0,55
	Tipo de instituição do ensino médio	Particular	VET-DPA	-0,652	0,52
	Qual é a faixa de renda mensal familiar?	Entre R$18,6mil e R$29,7mil	VQ5-DF4	-0,205	0,81
Sobre o vestibular	Você já prestou vestibular da FGV-EAESP?	Sim	VQ1	-0,355	0,70
	Meio de divulgação do vestibular	Internet	VQ8-D01	0,224	1,25

Fonte: elaboração própria

Por meio da análise do modelo de regressão logística, pode-se dizer que um aluno proveniente do interior de São Paulo, com bom conhecimento em matemática e ciências humanas, noções do idioma inglês, capaz de articular suas ideias através de um texto e teve conhecimento do vestibular por meio da internet pode ter um bom desempenho no curso de administração de empresas da FGV/EAESP. No entanto, suas chances diminuem se ele já prestou vestibular anteriormente, tiver mais de 18 anos quando ingressou, a renda mensal familiar for entre R$18,6mil e R$29,7mil e se tiver estudado em uma instituição particular no ensino médio.

Classificadores Bayesianos

Quatro classificadores Bayesianos foram desenvolvidos: Naïve Bayes, Tree augmented Naïve Bayes (TAN), BN Augmented Naïve Bayes (BAN) e um Naïve Bayes a partir das variáveis selecionadas no modelo de regressão logística.

Para os três primeiros classificadores Bayesianos, as 47 dummies definidas no passo 04 foram pré-selecionadas utilizando o índice de DKL, ou seja, a partir do ganho de informação causado pela variável dummy à variável alvo.

A Tabela 20 apresenta o índice de DKL das 47 potenciais variáveis dummies previsoras, que estão em ordem decrescente ao ganho de informação adicionado à variável alvo. Pelo resultado, a variável referente a faixa etária entre 19 e 20 anos do aluno ao ingressar no curso de administração é aquela com maior contribuição, 8,11%. Todas as outras possuem uma contribuição entre 7,7% e 0%, indicando que não há um conjunto de variáveis predominantes. Além do índice de DKL, foi feito o teste de independência de Chi^2 entre o par variável previsora e variável alvo. Com isso, foram eliminadas 22 variáveis, que tinham uma contribuição total de 5%. Consequentemente os classificadores foram construídos utilizando as 25 variáveis dummies previsoras selecionadas. A eliminação de 22 variáveis reduziu quase pela metade o número de variáveis previsoras em contrapartida a capacidade de discriminação dos classificadores Bayesianos não foi fortemente impactada porque a variação do índice ROC, foi no máximo de 2,2 pontos percentuais conforme Tabela 21. Como a redução de complexidade é significativa e a perda na capacidade de discriminação é mínima, a seleção de variáveis foi mantida.

Tabela 20 – Seleção de variáveis baseado no índice de DKL e teste de Chi^2

Descrição Variável Dummy	Código Variável	Índice de DKL	Contribuição da variável	Contribuição acumulada	Teste Chi²	P-Valor
Idade entre 19 e 20	VIE-D19a20	0,0175	8,1%	8,1%	23,25	0,00%
MatemáticaF1/Nota entre 7 e 8	VP1-D7a8	0,0167	7,7%	15,8%	22,11	0,00%
MatemáticaF1/Nota entre 8 e 9	VP1-D8a9	0,0146	6,7%	22,6%	19,32	0,00%
MatemáticaF2/Nota entre 7 e 8	VP5-D7a8	0,0132	6,1%	28,7%	17,45	0,00%
CienciasHumF1/Nota entre 7 e 8	VP4-D7a8	0,0131	6,0%	34,7%	17,28	0,00%
MatemáticaF1/Nota entre 6 e 7	VP1-D6a7	0,0119	5,5%	40,2%	15,71	0,01%
MatemáticaF2/Nota entre 6 e 7	VP5-D6a7	0,0112	5,2%	45,4%	14,82	0,01%
Idade entre 18 e 19	VIE-D18a19	0,0106	4,9%	50,2%	13,93	0,02%
MatemáticaF2/Nota acima de8	VP5-DA8	0,0102	4,7%	55,0%	13,40	0,03%
MatemáticaF2/Nota entre 5 e 6	VP5-D5a6	0,0099	4,6%	59,5%	13,10	0,03%
InglêsF1/Nota entre 5 e 6	VP3-D5a6	0,0091	4,2%	63,7%	11,99	0,05%
MatemáticaF1/Nota acima de 9	VP1-DA9	0,0088	4,0%	67,8%	11,53	0,07%
MatemáticaF1/Nota entre 5 e 6	VP1-D5a6	0,0063	2,9%	70,7%	8,24	0,41%
Idade acima de 20	VIE-DMaior20	0,0061	2,8%	73,5%	8,05	0,45%
RedaçãoF2/Nota acima de 6	VP6-DA6	0,0058	2,7%	76,2%	7,75	0,54%
InglêsF1/Nota entre 6 e 7	VP3-D6a7	0,0056	2,6%	78,8%	7,50	0,62%
InglêsF1/Nota acima de 7	VP3-DA7	0,0053	2,5%	81,3%	7,09	0,77%
Já prestou vestibular na FGV	VQ1	0,0046	2,1%	83,4%	6,12	1,34%
PortuguêsF1/Nota entre 6 e 7	VP2-D6a7	0,0044	2,0%	85,5%	5,89	1,52%
CienciasHumF1/Nota acima de 8	VP4-DA8	0,0044	2,0%	87,5%	5,81	1,59%
PortuguêsF1/Nota entre 7 e 8	VP2-D7a8	0,0043	2,0%	89,5%	5,75	1,65%
PortuguêsF1/Nota entre 5 e 6	VP2-D5a6	0,0039	1,8%	91,3%	5,20	2,25%
RedaçãoF2/Nota entre 5 e 6	VP6-D5a6	0,0025	1,2%	92,5%	3,39	6,57%
Região SP interior	VRE-DSPI	0,0023	1,1%	93,6%	3,02	8,23%
PortuguêsF1/Nota acima de 8	VP2-DA8	0,0022	1,0%	94,6%	2,86	9,06%

Continua

Descrição Variável Dummy	Código Variável	Índice de DKL	Contribuição da variável	Contribuição acumulada	Teste Chi²	P-Valor
PortuguêsF1/Nota entre 4 e 5	VP2-D4a5	0,0013	0,6%	95,2%	1,70	19,19%
Anuncio Vestibular (Feiras e Eventos)	VQ8-D07	0,0013	0,6%	95,7%	1,68	19,49%
Grau2 Particular	VET-DPA	0,0011	0,5%	96,3%	1,47	22,49%
Região Grande SP	VRE-DGSP	0,0011	0,5%	96,7%	1,40	23,76%
Candidato Muito Decidido	VQ4-DMD	0,0011	0,5%	97,2%	1,39	23,77%
Anúncio Vestibular (Outros)	VQ8-D03	0,0010	0,5%	97,7%	1,31	25,29%
Anúncio Vestibular (Internet)	VQ8-D01	0,0007	0,3%	98,0%	0,92	33,65%
Anúncio Vestibular (Cursinho, Colégio, Amigos e Familiares)	VQ8-D08	0,0006	0,3%	98,3%	0,80	37,26%
Anúncio Vestibular (Jornal e Revista)	VQ8-D04	0,0006	0,3%	98,6%	0,76	38,30%
R. Mensal De R$19k a R$28k	VQ5-DF4	0,0006	0,3%	98,8%	0,75	38,76%
Anuncio Vestibular (Não viu)	VQ8-D02	0,0005	0,2%	99,0%	0,66	41,73%
R. Mensal De R$9k a R$19k	VQ5-DF3	0,0004	0,2%	99,2%	0,56	45,43%
R. Mensal De R$28k a R$37k	VQ5-DF5	0,0004	0,2%	99,4%	0,47	49,39%
CienciasHumF1/Nota entre 4 e 7	VP4-D4a7	0,0003	0,1%	99,5%	0,40	52,68%
Nível Instrução Pai (Mestrado Doutorado)	VQ7-DEMD	0,0003	0,1%	99,7%	0,39	53,23%
Nível Instrução Pai (Mestrado Doutorado)	VQ6-DMD	0,0002	0,1%	99,8%	0,24	62,49%
Nível Instrução Pai (Sup. Completo)	VQ6-DSP	0,0001	0,1%	99,8%	0,17	68,02%
Anúncio Vestibular (Cinemas e TV)	VQ8-D06	0,0001	0,1%	99,9%	0,17	68,09%
R. Mensal acima de R$37k	VQ5-DF6	0,0001	0,0%	99,9%	0,11	73,49%
Candidato Decidido	VQ4-DDE	0,0001	0,0%	100,0%	0,09	76,36%
Nível Instrução Pai (Pós Grad)	VQ6-DPO	0,0001	0,0%	100,0%	0,08	78,20%
Nível Instrução Mae (Sup. Completo e Pós)	VQ7-DGSP	0,0000	0,0%	100,0%	0,06	81,21%

Conclusão

Fonte: elaboração própria

Tabela 21 - Comparativo do índice ROC entre classificadores Bayesianos com todas variáveis previsoras e classificadores com as variáveis previsoras selecionadas

Tipo de Base	Classificador Bayesiano	ROC - 47 variáveis previsoras	ROC - 22 variáveis previsoras	Variação em p.p.[4]
Base Desenvolvimento	Naïve Bayes	67,4%	66,7%	0,7%
	TAN	70,7%	68,6%	2,1%
	BAN	71,0%	68,8%	2,2%
Base Validação	Naïve Bayes	70,2%	69,5%	0,7%
	TAN	70,4%	68,7%	1,8%
	BAN	68,7%	67,9%	0,8%

Fonte: elaboração própria

Os três classificadores Bayesianos construídos a partir das 25 variáveis dummies selecionadas resultaram em modelos cujos índices ROC, TAB, TAM e TAT estão na Tabela 22. Analisando o índice ROC dos três modelos, todos os classificadores possuem um baixo

[4] p.p. é a abreviatura usa para a expressão pontos percentuais

poder de discriminação, conforme referências da Tabela 6. Além disso, ao analisar os classificadores estimados a partir da base de desenvolvimento, nota-se que o Naïve Bayes apresenta menor ROC, indicando que as variáveis previsoras não são condicionalmente independentes entre si. No entanto, o ganho obtido ao adicionar a dependência entre as variáveis é de 2 pontos percentuais porque o ROC sai de 66,7% para 68,6% a 68,8% nos classificadores TAN e BAN respectivamente. Essa melhora é marginal e aparentemente o aumento de complexidade no modelo não acarretou numa melhora significativa no seu poder de discriminação. Além disso, se fosse escolhido o classificador BAN por apresentar um ROC ligeiramente melhor, constata-se que ele possui *overfitting* à base de desenvolvimento porque o ROC na base de validação é 67,9%, 1 ponto percentual menor do que o seu ROC na base de desenvolvimento. O classificador TAN possui resultado semelhante ao BAN com a vantagem de aparentemente não ter *overffiting* à base de validação

Além do ROC, avaliam-se o TAB, o TAM e o TAT. Começando pela taxa de acerto de BOM, TAB, verifica-se que o aumento da complexidade do classificador não resulta em uma melhora significativa nesse índice. Veja na Tabela 22, que o TAB na base de desenvolvimento do classificador Naïve Bayes é de 64,9% e os respectivos TAB do classificador TAN e do classificador BAN são 65,5% e 65,9%. Com isso, ratifica o ganho marginal obtido com o aumento da complexidade. O TAM também apresenta comportamento semelhante ao TAB, por ele varia de 59,3% a 60,9%, podendo dizer que estatisticamente não possuem diferença. Analogamente conclui-se que o TAT também não teve melhora significativa com o aumento da complexidade no modelo.

Ao analisar os resultados dos classificadores na base de validação, verifica-se que o Naïve Bayes possui o maior ROC, 69,5% e o BAN o menor, 67,9%. O TAB é praticamente o mesmo entre eles e o classificador Naïve Bayes possui os maiores TAM e TAT. Sendo assim, o Naïve Bayes, Figura 17, é o classificador selecionado por apresentar menor complexidade, poder de discriminação semelhante aos outros classificadores Bayesianos na base de desenvolvimento e um resultado ligeiramente melhor na base de validação.

Tabela 22 – Índice ROC, TAB, TAM e TAT dos 3 classificadores Bayesianos

Índices	Naïve Bayes		TAN		BAN	
	Desenv.	Valid.	Desenv.	Valid.	Desenv.	Valid.
ROC	66,7%	69,5%	68,6%	68,7%	68,8%	67,9%
TAB[1]	64,9%	62,6%	65,5%	62,6%	65,9%	63,6%
TAM[1]	59,3%	68,3%	60,9%	66,8%	60,9%	63,4%
TAT[1]	62,3%	65,5%	63,3%	64,7%	63,5%	63,5%

1) TAB é taxa acerto BOM, TAB é taxa acerto MAU e TAT é taxa acerto Total

Fonte: elaboração própria

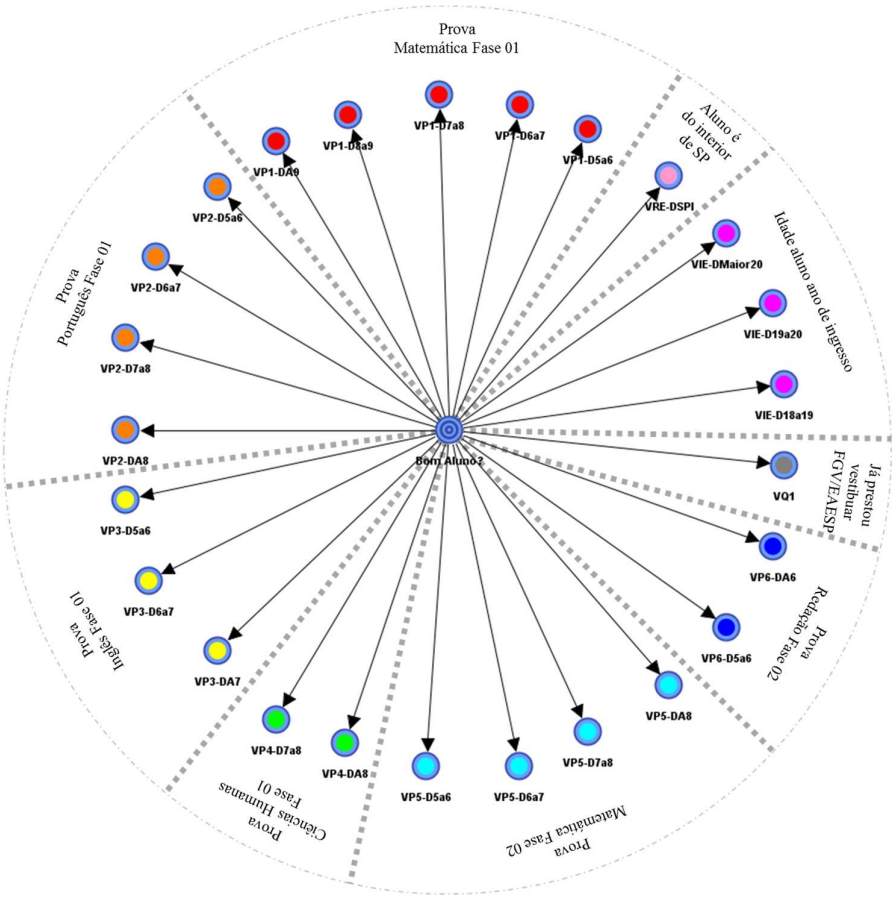

Figura 17 – Classificador Naïve Bayes com 25variáveis previsoras

Fonte: elaboração própria

O classificador Naïve Bayes é composto por variáveis previsoras que capturam o desempenho do aluno nas seis provas do vestibular, a sua idade no ano de ingresso, a região do Brasil de onde é proveniente e se o aluno realizou o vestibular da FGV/EAESP mais de uma vez antes de ser aprovado. Esse modelo indica que o grau de instrução dos pais, a renda mensal familiar, o tipo de instituição de ensino médio, o grau de decisão do aluno em estudar na

FGV/EAESP e os meios de comunicação para divulgação do vestibular não influenciam no desempenho do aluno.

Além desses três classificadores descritos, foi desenvolvido um classificador Naïve Bayes, denominado Naïve Bayes Reg.Log., a partir das 14 variáveis previsoras selecionadas no modelo de regressão logística binária. O comparativo da Tabela 23 entre os índices ROC, TAB, TAM e TAT do classificador Naïve Bayes e do Naïve Bayes Reg.Log. mostra que eles possuem diferenças pouco significativas. Tanto na base de desenvolvimento quanto na base de validação os índices ROC, TAB e TAT praticamente não apresentam diferença significativa. Apenas o TAM na base de validação apresenta uma variação maior a favor do classificador Naïve Bayes da Figura 17. Com isso não se pode afirmar que o critério de seleção de variáveis gerou um classificador melhor ou pior, porque ambos apresentam baixa capacidade de discriminação e taxas de acerto semelhantes.

Tabela 23 – Comparativo entre os classificadores Naïve Bayes e o Naïve Bayes Reg.Log.

Base	Comparativo	ROC	TAB[1]	TAM[1]	TAT[1]
Desenvolvimento	(1) Naïve Bayes	66,7%	64,9%	59,3%	62,3%
	(2) Naïve Bayes Reg. Log.	67,9%	66,7%	58,5%	62,8%
	(1)-(2) Variação	-1,2%	-1,8%	0,9%	-0,5%
Validação	(1) Naïve Bayes	69,5%	62,6%	68,3%	65,5%
	(2) Naïve Bayes Reg. Log.	69,3%	63,6%	63,9%	63,7%
	(1)-(2) Variação	0,2%	-1,0%	4,4%	1,7%

1) TAB é taxa acerto BOM, TAB é taxa acerto MAU e TAT é taxa acerto Total

Fonte: elaboração própria

Utilizando esses dois classificadores Bayesianos será analisada a influência das variáveis previsoras na variável alvo. A medida utilizada é a probabilidade de o aluno ter bom desempenho quando uma variável dummy é instanciada. A análise foca na tendência que as variáveis podem indicar sobre o desempenho do aluno.

Iniciando pelo classificador Naïve Bayes da Figura 17, o Gráfico 1 indica que a probabilidade de o aluno ter bom desempenho é inversamente proporcional à idade do aluno no ano de ingresso. Alunos provenientes do interior de São Paulo tem uma maior probabilidade de ter bom desempenho, conforme Gráfico 2. A quantidade de vezes que o aluno prestou vestibular na FGV/EAESP é inversamente proporcional à probabilidade de o aluno ter bom desempenho, ou seja, um aluno aprovado na primeira vez que prestou vestibular tem uma maior probabilidade de ter bom desempenho do que um aluno que prestou o vestibular mais de uma vez antes de ser aprovado na FGV/EAESP.

Os gráficos das notas nas provas do vestibular indicam que a probabilidade de um aluno ter bom desempenho é diretamente proporcional à nota das provas. Aparentemente o desempenho nas provas de matemática geram uma maior variação na probabilidade de o aluno ter bom desempenho. Isso é observado no Gráfico 4 porque a probabilidade de bom desempenho sobe de 20%, se a nota for entre 5 e 6, para quase 80%, se a nota for acima de 8, e no Gráfico 8 porque a probabilidade sobe de 37%, se a nota for entre 5 e 6, para 76% se a nota for acima de 8. A nota da prova de português da primeira fase apresenta uma variação menor na probabilidade de o aluno ter bom desempenho porque sobe de 45% para 70%. A nota da prova de inglês indica que alunos com média acima de 6 tem uma probabilidade acima de 50% de ser bom aluno, Gráfico 6, assim como a nota da prova de redação da fase 02, Gráfico 9. O Gráfico 7 mostra que alunos com nota acima de 7 na prova de ciências humanas tem uma probabilidade acima de 60% de ter bom desempenho.

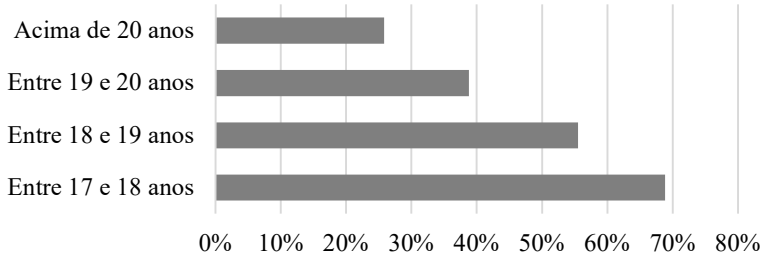

Gráfico 1 – Probabilidade de aluno ter bom desempenho dependendo da idade do aluno no ano de ingresso – classificador Naïve Bayes
Fonte: elaboração própria

Gráfico 2 – Probabilidade de aluno ter bom desempenho dependendo da região do Brasil de onde o aluno é proveniente – classificador Naïve Bayes
Fonte: elaboração própria

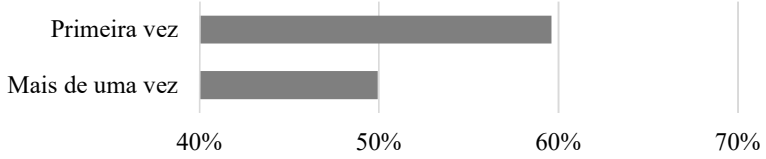

Gráfico 3 – Probabilidade de aluno ter bom desempenho dependendo da quantidade de vezes que o aluno prestou vestibular na FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

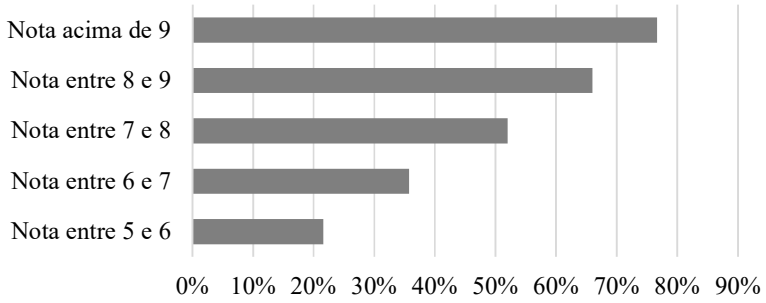

Gráfico 4 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de matemática na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

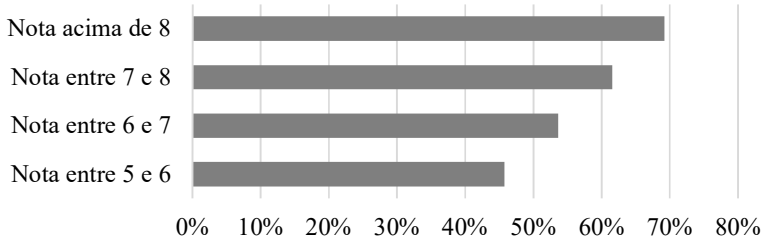

Gráfico 5 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de português na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

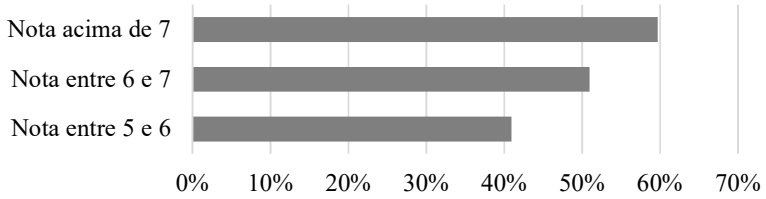

Gráfico 6 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de inglês na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

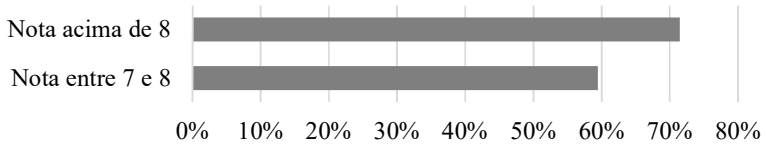

Gráfico 7 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de ciências humanas na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

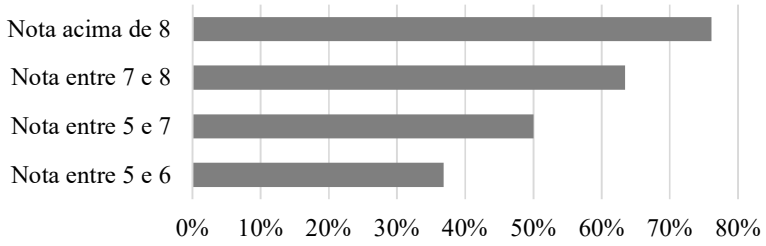

Gráfico 8 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de matemática na fase 02 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

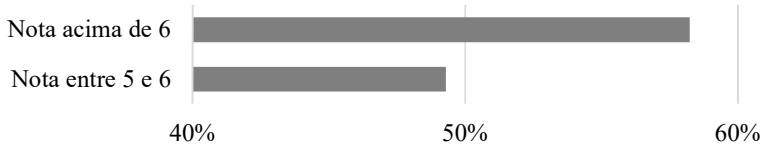

Gráfico 9 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de redação na Fase 02 do vestibular da FGV/EAESP – classificador Naïve Bayes
Fonte: elaboração própria

Após essa análise, o classificador Naïve Bayes estima que alunos com maior probabilidade de bom desempenho ingressam na FGV/EAESP com idade entre 17 e 18 anos, vieram do interior de São Paulo e realizaram o vestibular apenas uma vez. Em termos de desempenho nas provas, boas notas nas provas de matemática aumenta a probabilidade de bom desempenho, assim como ter conhecimento do idioma inglês e português. Conhecer sobre ciências humanas e ter capacidade de articular suas ideias através de um texto também aumentam a probabilidade de o aluno ter bom desempenho.

A mesma análise é feita para o classificador Naïve Bayes Reg. Log. com os resultados apresentados nos gráficos 10 a 20. A idade do aluno no ano de ingresso é inversamente proporcional à sua probabilidade de bom desempenho, como indica o Gráfico 10. Alunos provenientes do interior de São Paulo possuem uma maior probabilidade de ter bom desempenho e o número de vezes que o aluno já prestou vestibular da FGV/EAESP é inversamente proporcional à probabilidade de ele ter bom desempenho, Gráfico 11. O desempenho das notas nas provas do vestibular é diretamente proporcional à probabilidade de o aluno ter bom desempenho, como indicam os gráficos de 12 a 17.

A renda mensal da família entre R$18,6 mil e R$29,7 mil e a internet como meio de comunicação para obter conhecimento sobre o vestibular da FGV/EAESP influenciam a probabilidade de bom desempenho. No entanto, o Gráfico 18 indica que se a família tiver essa

renda mensal, a probabilidade reduz de 53,6% para 50,8% e o Gráfico 19 indica que a internet como meio de divulgação do vestibular aumenta a probabilidade de 50,8% para 54,0%, mostrando que essas variáveis previsoras exercem pouca influência na variável alvo, como observado por meio do índice DKL. O tipo de instituição do ensino médio aumenta as chances de o aluno ter bom desempenho, conforme Gráfico 20.

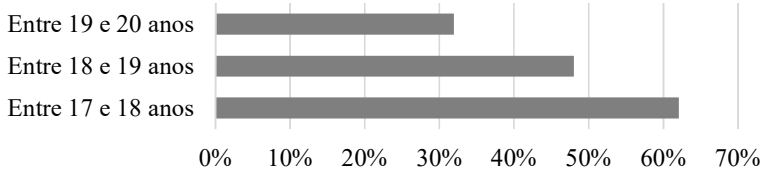

Gráfico 10 – Probabilidade de aluno ter bom desempenho dependendo da idade do aluno no ano de ingresso – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

Gráfico 11 – Probabilidade de aluno ter bom desempenho dependendo da região do Brasil de onde o aluno é proveniente – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

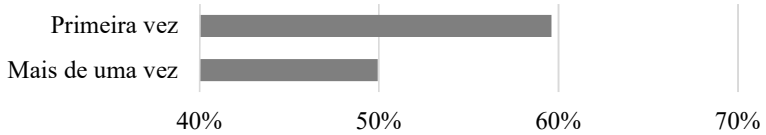

Gráfico 12 – Probabilidade de aluno ter bom desempenho dependendo da quantidade de vezes que o aluno prestou vestibular na FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

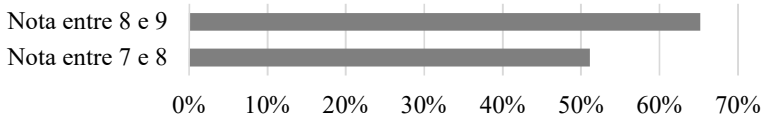

Gráfico 13 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de matemática na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

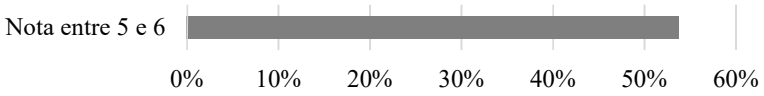

Gráfico 14 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de inglês na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

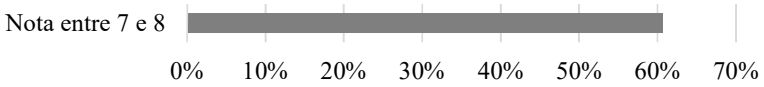

Gráfico 15 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de ciências humanas na fase 01 do vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

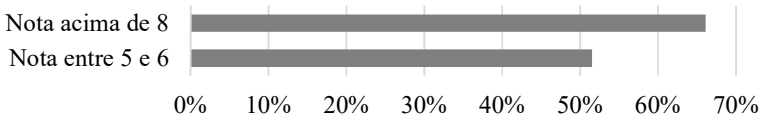

Gráfico 16 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de matemática na fase 02 do vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

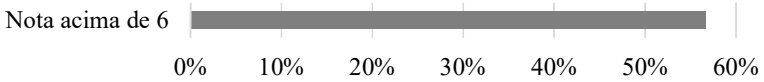

Gráfico 17 – Probabilidade de aluno ter bom desempenho dependendo da nota da prova de redação na fase 02 do vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

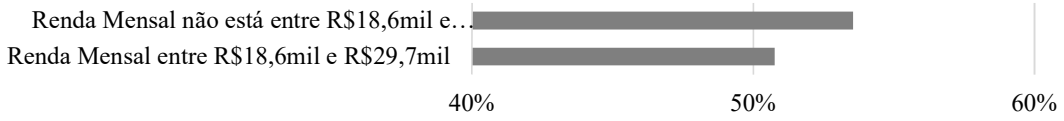

Gráfico 18 – Probabilidade de aluno ter bom desempenho dependendo da renda mensal familiar – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

Gráfico 19 – Probabilidade de aluno ter bom desempenho dependendo do meio de comunicação para divulgar o vestibular da FGV/EAESP – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

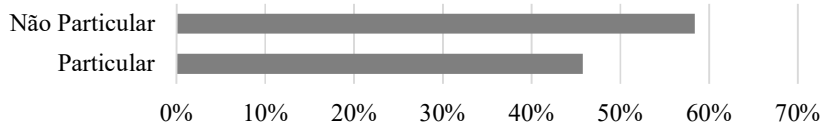

Gráfico 20 – Probabilidade de aluno ter bom desempenho dependendo do tipo de instituição de ensino médio – classificador Naïve Bayes Reg. Log.
Fonte: elaboração própria

Como os dois classificadores indicam conclusões semelhantes e o classificador Naïve Bayes Reg. Log. apresenta as discrepâncias devido as variáveis previsoras de renda mensal, da internet como meio de comunicação e do tipo de instituição de ensino médio, o classificador Naïve Bayes com 25 variáveis previsoras da Figura 17 será utilizado para comparação com o modelo de regressão logística.

Comparativo entre regressão logística e classificador Bayesiano

Após estimar o modelo de regressão logística obtido pelo critério de seleção de variáveis *Stepwise* e o classificador Naïve Bayes com 25 variáveis, eles serão comparados para avaliar se apresentam resultados muito diferentes.

Conforme indica a Tabela 24, na base de desenvolvimento, o modelo de regressão logística apresenta um ROC 2,3 pontos percentuais melhor do que o ROC do classificador Naïve Bayes. Na base de validação ambos apresentam o mesmo ROC. No entanto, ambos possuem baixa capacidade de discriminação conforme critério detalhado na Tabela 6. A taxa de acerto de BOM é melhor no modelo de regressão logística nas duas bases de dados e o classificador Naïve Bayes apresenta TAM melhor. Embora exista essas pequenas diferenças, ambos modelos tem uma taxa de acerto total, TAT, em torno de entre 62% e 65%, ou seja, 35% a 38% dos casos seriam classificados incorretamente.

Tabela 24 – Comparativo entre os índices ROC, TAB, TAM e TAT dos modelos de regressão logística e classificador Naïve Bayes

Índices	Regressão logística - Stepwise		Classificador Naïve Bayes	
	Desenv.	Valid.	Desenv.	Valid.
ROC	69,0%	69,0%	66,7%	69,5%
TAB[1]	68,5%	64,1%	64,9%	62,6%
TAM[1]	58,7%	63,4%	59,3%	68,3%
TAT[1]	63,8%	63,7%	62,3%	65,5%

1) TAB é taxa acerto BOM, TAB é taxa acerto MAU e TAT é taxa acerto Total

Fonte: elaboração própria

Apesar de as variáveis previsoras não serem as mesmas, as notas das provas de matemática, inglês e ciências humanas da fase 01 estão presentes em ambos modelos, conforme Tabela 25. A idade no ano de ingresso, a região de onde o aluno é proveniente e a quantidade de vezes que o aluno prestou o vestibular da FGV/EAESP também estão presentes em ambos modelos. Note que as variáveis em comum possuem o mesmo tipo de influência na probabilidade de o aluno ter bom desempenho corroborando para a consistência dos resultados. A nota da prova de português da fase 01 está presente apenas no classificador Naïve Bayes. Já a renda mensal, a internet como meio de divulgação do vestibular da FGV/EAESP e o tipo de instituição de ensino médio influenciam apenas o modelo de regressão logística.

As variáveis previsoras presentes em apenas um dos modelos são as divergências entre eles e há razões para mantê-las. A nota da prova de português da fase 01 influencia na variável alvo de acordo com o índice de DKL e as outras três variáveis presentes apenas no modelo de regressão logística exercem baixa influência na variável alvo de acordo com o índice DKL. Ao analisar o teste de independência de Chi^2, não pode dizer que a variável alvo é dependente dessas três variáveis previsoras, eliminando-as. Ao olhar na perspectiva do modelo de regressão logística, a nota da prova de português da fase 01 não é significante para o modelo, em contrapartida a renda mensal, a internet como meio de divulgação e o tipo de instituição são significantes. Embora exista argumento para mantê-las nos respectivos modelos, baseando-se nos resultados destes dois modelos não há consenso sobre elas.

Em termos de número de variáveis previsoras, o modelo de regressão logística possui 14 e o classificador Naïve Bayes 25, indicando que o primeiro modelo apresenta menor complexidade.

Baseando-se nas análises anteriores deste trabalho, pode-se dizer que a probabilidade de um aluno ter bom desempenho é inversamente proporcional à idade do aluno no ano de ingresso e ao número de vezes que ele prestou o vestibular da FGV/EAESP. Além

disso, alunos com boa probabilidade de ter bom desempenho vieram do interior de São Paulo, tiveram boas notas nas provas de matemática, possuem conhecimento no idioma inglês, apresentam bom conhecimento em ciências humanas e são capazes de articular suas ideias por meio de um texto.

Tabela 25 – Grupo de variáveis previsoras presentes nos classificadores

↓ Diminui probabilidade ↑ Aumenta probabilidade

Grupo das variáveis previsoras	Tema das variáveis previsoras	Presente em:		
		Somente Regressão Logística	Somente Naïve Bayes	Ambos modelos
Provas do Vestibular	Matemática - Fase 01			↑
	Português - Fase 01		↑	
	Inglês - Fase 01			↑
	Ciências Humanas - Fase 01			↑
	Matemática Aplicada - Fase 02			↑
	Redação - Fase 02			↑
Informações Socio-Econômicas	Região onde morava durante ensino médio (SP interior)			↑
	Idade do aluno no ano de ingresso			↓
	Tipo de instituição do ensino médio	↓		
	Renda mensal familiar (R$18,6mil a R$29,7mil)	↓		
Vestibular	Você já prestou vestibular da FGV-EAESP?			↓
	Meios de divulgação - Internet	↑		

Fonte: elaboração própria

5 Conclusão

Este trabalho desenvolveu classificadores utilizando modelo de regressão logística binária e redes Bayesianas a partir de uma base de dados completa. A metodologia apresentada, derivada do processo de KDD desenvolvido por Fayyad et al. (1996a), estruturou os passos a serem seguidos para a mineração de dados apresentada.

Nas bases de dados coletadas foram identificadas 187 informações para cada registro da amostra e a partir delas geradas 16 potenciais variáveis previsoras que produziram 47 variáveis dummies. A variável alvo indica se o aluno tem bom desempenho ou não. Para definir desempenho foram consideradas as disciplinas obrigatórias cursadas durante os quatro semestres iniciais do curso de administração da FGV/EAESP. Alunos com até uma DP é considerado com bom desempenho. Essas potenciais variáveis previsoras e a variável alvo foram usadas para estimar classificadores utilizando regressão logística e classificadores Bayesianos.

O modelo de regressão logística foi estimado utilizando três critérios de seleção de variáveis: Seleção *Forward,* Eliminação *Backward* e *Stepwise*. Para testar a aderência do modelo à amostra foi utilizado o teste de Hosmer e Lemeshow. As variáveis selecionadas são referentes às provas do vestibular exceto a prova de português, a idade do aluno no ano de ingresso, a região do Brasil de onde o aluno é proveniente, quantas vezes ele prestou vestibular antes de ser aprovado, o tipo de instituição do ensino médio, a renda familiar mensal e a internet como meio de divulgação do processo de vestibular.

Para os três modelos de regressão logística foram calculados os índices de ROC, TAB, TAT e TAM. Ao analisa-los, verifica-se que eles possuem aderência à amostra e apresentam ROC em torno de 69%, indicando baixo poder de discriminação. Dentre os três modelos de regressão logística foi escolhido o modelo baseado no critério *Stepwise* por ter a menor quantidade de variáveis previsoras e consequentemente menor complexidade.

Embora a capacidade de discriminação do modelo seja baixa, as variáveis previsoras foram analisadas e sinalizam que a probabilidade de um aluno ter bom desempenho é inversamente proporcional à idade dele no ano de ingresso e ao número de vezes que ele prestou o vestibular da FGV/EAESP antes de ser aprovado. Além disso, as variáveis previsoras indicam que alunos que vieram do interior de São Paulo possuem maior probabilidade de ter um bom desempenho. Em relação as notas das provas do vestibular, o modelo de regressão logística indicou que alunos com bom desempenho em matemática, conhecimentos em inglês e em ciências humanas e capazes de estruturar suas ideias em um texto têm uma maior

probabilidade de ter bom desempenho. O modelo de regressão logística também mostrou que alunos com renda mensal entre R$18,6 mil e R$29,7 mil tem menos probabilidade de ter bom desempenho, assim como aqueles que cursaram o ensino médio em uma instituição particular. Por fim, o modelo indica que alunos que conheceram o vestibular da FGV/EAESP pela internet tem maior probabilidade de ter bom desempenho.

Os quatro classificadores Bayesianos também foram construídos a partir das 47 variáveis dummies e da variável alvo. Para selecionar as variáveis previsoras foram analisados o ganho de informação que cada uma delas adiciona à variável alvo e o teste de independência de Chi^2 entre cada variável previsora e a variável alvo. As variáveis previsoras selecionadas foram aquelas relacionadas às seis provas do vestibular, a idade do aluno no ano de ingresso, a região do Brasil de onde ele vem e a quantidade de vezes que prestou vestibular antes de ser aprovado. Como resultado, foram mantidas 25 variáveis dummies utilizadas para estimar três classificadores Bayesianos: Naïve Bayes, Tree Augmented Naïve Bayes (TAN) e o Augmented Naïve Bayes (BAN).

O índice ROC dos três classificadores está entre 67% e 69% indicando que eles possuem baixo poder de discriminação. O classificador Naïve Bayes apresentou o menor ROC, 67%, no entanto, a adição de dependência entre as variáveis previsoras como foi feito no classificador TAN e no BAN adicionou apenas 2 pontos percentuais no ROC. Esse aumento de complexidade no modelo não é justificado pelo ganho obtido, os três classificadores possuem índices TAB, TAM e TAT praticamente iguais, sendo em torno de 65%, 60% e 63% respectivamente. Sendo assim, foi selecionando o classificador Naïve Bayes por apresentar menor complexidade e capacidade de discriminação semelhante aos outros dois classificadores Bayesianos.

O classificador Naïve Bayes também indica que a probabilidade de um aluno ter bom desempenho é inversamente proporcional à sua idade no ano de ingresso e a quantidade de vezes que prestou o vestibular antes de ser aprovado. Alunos com bom conhecimento de matemática, inglês, ciências humanas e português e com capacidade de estruturar suas ideias por meio de textos têm maior probabilidade de ter bom desempenho.

Um outro classificador Bayesiano foi estimado a partir do critério de seleção de variável *Stepwise* e estrutura de rede Naïve Bayes. O resultado foi o classificador Naïve Bayes Reg. Log. com um ROC 68%, muito próximo ao valor do outro classificador Naïve Bayes, indicando que o critério de seleção de variáveis exerceu pouca influência no resultado dos classificadores.

Ao comparar os resultados do modelo de regressão logística e do classificador Naïve Bayes, nota-se que os dois apresentam baixo poder de discriminação, sendo uma das contribuições deste estudo. Além disso, as variáveis previsoras comuns aos dois modelos, influenciam da mesma maneira na probabilidade de o aluno ter bom desempenho. Nos dois modelos essa probabilidade é diretamente proporcional às variáveis previsoras referentes à idade do aluno no ano de ingresso, região do Brasil de onde o aluno é proveniente, quantidade de vezes que o aluno prestou vestibular antes de ser aprovado e nota das provas de matemática da fase 01 e fase 02, inglês, ciências humanas e redação. Os dois modelos também indicam que essa probabilidade é inversamente proporcional à idade do aluno no ano de ingresso e a quantidade de vezes que o aluno prestou vestibular. Essa probabilidade aumenta se o aluno é proveniente do interior de São Paulo. Aparentemente nem o grau de formação dos pais nem o nível de decisão em estudar na FGV/EAESP influenciam na probabilidade. Essa coerência sinaliza quais informações podem indicar que um aluno terá bom desempenho no curso.

A identificação dessas informações como variáveis previsoras é outra contribuição do estudo, principalmente para pessoas interessadas em conhecer quais informações podem sinalizar que um aluno terá bom desempenho no curso de graduação de administração de empresas nas FGV/EAESP. A influência da nota da prova de português da fase 01, a renda familiar mensal, o tipo de instituição do ensino médio e a internet como meio de comunicação para divulgar o vestibular da FGV/EAESP são variáveis previsoras que não estão presentes nos dois modelos, não sendo consenso a influência delas sobre a variável alvo.

Embora os classificadores possuem baixo poder de discriminação, eles podem ser utilizados para entender a influência das variáveis previsoras na variável alvo. Com isso, a primeira hipótese do estudo foi confirmada, ou seja, as informações coletadas no processo de vestibular permitem estimar classificadores Bayesianos e modelos de regressão logística. A segunda hipótese do estudo também foi confirmada porque foi possível identificar as informações que sinalizam se o aluno terá um bom desempenho na graduação.

Como futuros estudos, para tentar aumentar o poder de discriminação do classificador Bayesiano, poderia usar uma abordagem mista para estima-lo. Para isso, o pesquisador pode utilizar a mesma base de dados e incluir o conhecimento de especialistas em educação para analisar a relação entre as variáveis. Isso pode esclarecer se a renda mensal, a nota da prova de português da fase 01 e o tipo de instituição de ensino influenciam ou não. Outra modificação que pode ser feita, é incluir novas informações a base de dados, por exemplo, atividades extracurriculares no ensino médio e na graduação.

Outra sugestão é realizar o mesmo estudo para identificar os atributos dos candidatos que possuem maior probabilidade de ser aprovado no vestibular no curso de administração de empresas da FGV/EAESP.

REFERÊNCIAS

CHENG, J., Greiner, R. Comparing Bayesian Network Classifiers. In: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI'99). Morgan Kaufmann, p.101-107, 1999.

CHOW, CK. AND LIU, C.N. Approximating discrete probability distributions with dependence trees. IEEE Trans. On Information Theory, 14 (pp. 462-467), 1968.

CONRADY, S., JOUFFE, L. Bayesian Networks & Bayesia Lab. A practical introduction for Researchers, 2015. Disponível em <http://www.bayesialab.com/book>, Acesso em 01 de setembro de 2015.

COVER, T.M, THOMAS, J. A. Elements of Information Theory. 2.ed. Hoboken, New Jersey: John Wiley & Sons, 2006.

CRESWELL, J.W. Projeto de pesquisa: Métodos qualitativo, quantitativo e misto; tradução Magda Lopes. 3.ed. Porto Alegre: Bookman, 2010.

FAYYAD, U., PIAIATETSKY-SHAPIRO, G., E SMYTH, P. From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining. Cambridge, Massachusetts: AAAI/MIT Press, 1996.

FAYYAD, U., PIATETSKY-SHAPIRO, G., E SMYTH, P. Knowledge Discovery and Data Mining: Towards a unifying framework. KDD V.96, p82-88. AAAI, 1996.

FAYYAD, U., PIATETSKY-SHAPIRO, G., E SMYTH, P. The KDD Process for extracting useful knowledge from volumes of data. Communications of the ACM, v.39, n. 11, p. 27-34, November 1996.

FRIEDMAN, N. GEIGER, D. GOLDSZMIDT, M. Bayesian Network Classifiers. Machine Learning, v.29, p. 131-163. The Netherlands: Kluwer Academic Publishers, 1997.

HECKERMAN, D., A Tutorial Learning With Bayesian Network, Microsoft Research. Technical Report, Advanced Technology Division, Microsoft Corporation. March 1995, Revised November 1996.

HOSMER, D.W., LEMESHOW, S., STURDIVANT, R.X. Applied logistic regression. 3.ed. Hoboken, New Jersey: Wiley, 2013.

JENSEN, F.V., An Introduction to Bayesian Networks. 1.ed. Denmak: Springer, 1996.

JENSEN, F.V., NIELSEN, T.D., Bayesian Networks and Decision Graphs, 2.ed. Springer, 2007.

LENGLEY, P., IBA, W., THOMPSON, K. An analysis of Bayesian Classifiers, In Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose: AAAI, 1992.

KARCHER, C. Redes Bayesianas Aplicadas à análise do risco de crédito. 2009. Dissertação Mestrado – Escola Politécnica da Universidade de São Paulo. Departamento de Sistemas Eletrônicos, São Paulo.

MAMLET, R., VANDEVELDE, C. College Admission: From application to acceptance, step by step. Updated Edition. New York, USA: Three River Press, 2011.

MORETTIN P.A, BUSSAB, W.O. Estatística Básica. 5.ed. São Paulo: Saraiva, 2004.

NEAPOLITAN R.E. Learning Bayesian Networks. Chicago, Illinois: Prentice Hall Series in Artificial Intelligence, 2003.

NOGUEIRA, C. M. Dificuldades orçamentárias básicas das famílias brasileiras: um convite à reflexão a partir de redes bayesianas. 2012. Dissertação Doutorado – Universidade de São Paulo, São Paulo.

TOMASELLA S.M.O., PEDROSO A.C. e SICSÚ A. L. Análise empírica dos indicadores KS e ROC. Relatório de pesquisa, Centro de Excelência Bancária, FGV – EAESP, São Paulo, 2008.

SIDDIQI, N. Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

SICSÚ, A.L. Credit scoring: desenvolvimento, implantação, acompanhamento. 1.ed. São Paulo: Blucher, 2010.

WEST, D.B. Introduction to Graph Theory. 2.ed. New Jersey: Prentice-Hall, 2000.
A divisão do sistema de educação brasileiro. Disponível em

<http://www.brasil.gov.br/educacao/2014/05/saiba-como-e-a-divisao-do-sistema-de-educacao-brasileiro/image_view_fullscreen>, Acesso em 20 de outubro de 2015.

WITTEN, I.H; FRANK E. Data Mining Practical Machine Learning Tools and Techniques. 2.ed. Elsevier, 2005.

APÊNDICE A - Lista de informação disponível nas bases de dados

		Bases de Dados								Análise	
Informações contidas nas bases de dados	**Descritivo da informação**	**a) Dados de Inscrição**	**b) Respostas do questionário**	**c) Notas do vestibular**	**a) Dados do vestibular**	**b) Dados dos alunos da graduação**	**c) Dados do perfil do aluno de graduação**	**d) Notas das disciplinas da graduação**	**Número de Repetições**	**Avaliação**
Nome	Nome do candidato/aluno	x	x	x	x	x	x	x	7	Caracterização
Ano de Ingresso	Ano de ingresso no curso de administração de empresa na FGV/EAESP	x	x	x	x	x	x	x	7	Caracterização
Semestre Ingresso	Semestre de ingresso no curso de administração de empresa na FGV/EAESP	x	x	x	x	x	x	x	7	Caracterização
Número de matrícula do aluno	Matrícula do aluno no curso				x	x	x	x	4	Caracterização
Curso	Nome do curso				x	x	x	x	4	Caracterização
Código do aluno	Código do aluno no curso				x		x	x	3	Caracterização
Tipo aluno	comum.				x	x	x		3	Caracterização
Situação	Situação do aluno na FGV/EAESP. Pode ser Ativo, quando está cursando, Concluído quando terminou o curso, Evadido, Jubilado quando é eliminado do curso de acordo com as regras da instituição, Cancelado quando abandono de curso, desistência do curso, cancelamento de matricula, novos ingressantes que não efetuaram matrícula, embora				x	x	x		3	Caracterização
CPF	Número do CPF do candidato/aluno	x		x			x		3	Caracterização
Currículo	Currículo vigente para o aluno. Ele define as disciplinas obrigatórias cursadas pelo aluno. Currículos com sufixo: _R / _Ac 1 / Ac 2 / reing / -Acerto podem ser considerados iguais aos currículos de mesmo nome até o sufixo, por exemplo CGAEN_1201RG1_R é igual a CGAEN_1201RG1					x	x		2	Caracterização
Sexo	Sexo do candidato/aluno	x					x		2	Caracterização
Data Nascimento	Data de nascimento do candidato/aluno	x					x		2	Potencial variável previsora
Nacionalidade	Nacionalidade do candidato/aluno	x					x		2	Caracterização
Estado Civil	Estado civil do candidato/aluno	x					x		2	Caracterização
RG	Número do RG do candidato/aluno	x					x		2	Caracterização
Órgão Emissor RG	Órgão responsável pela emissão do passaporte	x							1	Informação redundante
UF Emissor RG	Unidade Federativa onde o RG foi emitido	x							1	Informação redundante
Data Expedição RG	Data de expedição do RG	x							1	Informação redundante
Endereço	Endereço do candidato/aluno	x					x		2	Informação redundante
Complemento	Complemento do endereço do candidato/aluno	x					x		2	Informação redundante
Bairro	Bairro do endereço do candidato/aluno	x					x		2	Informação redundante
Cidade	Cidade do endereço do candidato/aluno	x					x		2	Informação redundante
Estado	Estado do endereço do candidato/aluno	x					x		2	Informação redundante
CEP	CEP do endereço do candidato/aluno	x					x		2	Potencial variável
E-mail pessoal	Correio eletrônico do candidato/aluno	x					x		2	Caracterização
Telefone 1	Número de telefone do candidato/aluno	x					x		2	Caracterização
Telefone 2	Número de telefone do candidato/aluno	x					x		2	Caracterização
Telefone 3	Número de telefone do candidato/aluno	x					x		2	Caracterização
Passaporte	Número do passaporte do aluno						x		1	Caracterização
RNE	Número da RNE do candidato/aluno estrangeiro	x							1	Caracterização
Validade RNE	Data de validade da RNE	x							1	Caracterização
E-mail comercial	Correio eletrônico comercial do candidato/aluno	x							1	Caracterização
Ex-aluno	Indica se o candidato é um ex-aluno	x		x					2	Caracterização
Número de Inscrição	Número de inscrição no vestibular	x	x	x					3	Caracterização
Candidato é treineiro?	Indica se o candidato é treineiro, ou seja, ainda não é formado no ensino médio, mas está realizando a prova para	x		x					2	Caracterização
Data de Inscrição	Data de inscrição no vestibular	x							1	Caracterização
Confirmado	Confirmação da inscrição	x							1	Caracterização
Primeira Opção	Indica qual curso é a primeira opção do candidato. Isso foi vigente até o primeiro semestre de 2011, quando eram dadas 2 opções de curso Administração de empresa e Administração pública	x		x					2	Caracterização
Segunda Opção	Indica qual curso é a segunda opção do candidato. Isso foi vigente até o primeiro semestre de 2011, quando eram dadas 2 opções de curso Administração de empresa e Administração pública	x		x					2	Caracterização
Matriculado	Indica se o candidato foi matriculado no curso	x		x					2	Caracterização
Ano Conclusão	Ano de conclusão do ensino médio	x							1	Caracterização
Ensino Médio Instituição	Nome da instituição do ensino médio	x					x		2	Informação redundante
Ensino Médio Unidade	Unidade do ensino médio quando o colégio possui mais de uma unidade	x							1	Informação redundante
Ensino Médio Cidade	Cidade onde o candidato realizou o ensino médio	x							1	Informação redundante
Ensino Médio Estado	Estado onde o candidato realizou o ensino médio	x							1	Informação redundante
Ensino Médio País	País onde o candito realizou o ensino médio	x							1	Informação redundante
Ensino Médio Período	Horário que o candidato estudou durante o ensino médio, podendo ser manhã, tarde e integral	x							1	Potencial variável
Ensino Médio Tipo Estab.	Tipo de estabelecimento do ensino médio podendo ser público ou privado	x							1	Potencial variável
Cursinho Instituição	Nome do curso pré-vestibular realizado pelo candidato	x							1	Não presente em todos
Cursinho Unidade	Unidade do curso pré-vestibular realizado pelo candidato	x							1	Não presente em todos
Cursinho Cidade	Cidade onde é localizado o curso pré-vestibular realizado pelo candidato	x							1	Não presente em todos
Cursinho Estado	Estado onde é localizado o curso pré-vestibular realizado pelo candidato	x							1	Não presente em todos
Cursinho País	País onde é localizado o curso-pré vestibular realizado pelo candidato	x							1	Não presente em todos
Cursinho Período	Horário em que o candidatou estudou no curso-pré vestibular	x							1	Não presente em todos
Cursinho Tipo Estab.	Topo de estabelecimento do curso pré-vestibular, podendo ser público ou privado	x							1	Não presente em todos
Primeiro Local	Local de preferência para realização das provas	x							1	Caracterização
Processo	Processo pelo qual o aluno ingressou na FGV/EAESP podendo ser vestibular, transferência, etc		x	x					2	Caracterização
Candidato	Número do candidato no vestibular			x					1	Caracterização
Identificação	Número de identificação no vestibular			x					1	Caracterização
Pos Clas 1	Posição na lista do aluno para a primeira opção de curso selecionado no momento da inscrição do vestibular			x					1	Multicolinearidade
Pos Clas 2	Posição na lista do aluno para a segunda opção de curso selecionado no momento da inscrição do vestibular			x					1	Multicolinearidade
Prova 1 bruta - Matemática Fase 01	Nota de 0 a 10 da prova de matemática Fase 01			x	x				2	Potencial variável
Prova 1 Padrão Vestibular	Nota relativa em relação aos candidatos da prova de matemática Fase 01			x	x				2	Multicolinearidade
Prova 2 bruta - Português Fase 01	Nota de 0 a 10 da prova de português Fase 01			x	x				2	Potencial variável
Prova 2 Padrão Vestibular	Nota relativa em relação aos candidatos da prova de português Fase 01			x	x				2	Multicolinearidade
Prova 3 bruta - Inglês Fase 01	Nota de 0 a 10 da prova de inglês Fase 01			x	x				2	Potencial variável
Prova 3 Padrão Vestibular	Nota relativa em relação aos candidatos da prova de inglês Fase 01			x	x				2	Multicolinearidade
Prova 4 bruta - Ciências Humanas Fase 01	Nota de 0 a 10 da prova de ciências humanas Fase 01			x	x				2	Potencial variável
Prova 4 Padrão Vestibular	Nota relativa em relação aos candidatos da prova de ciências humanas Fase 01			x	x				2	Multicolinearidade
Prova 5 bruta - Matemática Fase 02	Nota de 0 a 10 da prova de matemática aplicada Fase 02			x	x				2	Potencial variável
Prova 5 Padrão Vestibular	Nota relativa em relação aos candidatos da prova de matemática aplicada Fase 02			x	x				2	Multicolinearidade
Prova 6 bruta - Redação Fase 02	Nota de 0 a 10 da redação Fase 02			x	x				2	Potencial variável
Prova 6 Padrão Vestibular	Nota relativa em relação aos candidatos da redação Fase 02			x	x				2	Multicolinearidade

Continua

		Bases de Dados							Análise	
Informações contidas nas bases de dados	**Descritivo da informação**	**a) Dados de Inscrição**	**b) Respostas do questionário**	**c) Notas do vestibular**	**a) Dados do vestibular**	**b) Dados dos alunos da graduação**	**c) Dados do perfil do aluno de graduação**	**d) Notas das disciplinas da graduação**	**Número de Repetições**	**Avaliação**
Médi Fase1	Média das provas da primeira fase			x	x				2	Multicolinearidade
Clas Fase1	Classificação do candidato na primeira fase			x					1	Multicolinearidade
Status Fase1	Situação do candidato na primeira fase, podendo ser eliminado ou classificado			x					1	Caracterização
Médi Fase2	Média das provas da segunda fase			x	x				2	Multicolinearidade
Clas Fase2	Classificação do candidato na segunda fase			x					1	Multicolinearidade
Status Fase2	Situação do candidato na primeira fase, podendo ser eliminado ou classificado			x					1	Caracterização
Media Final	Média final do candidato no vestibular levando em consideração as notas da primeira e segunda fase			x					1	Classificação
Classificação Final	Classificação final do candidato			x	x				2	Multicolinearidade
Curso	Nome do curso						x		1	Caracterização
Empresa	Empresa onde o aluno trabalha						x		1	Caracterização
Cargo	Cargo atual do aluno						x		1	Caracterização
Ano Concl.Grad.	Ano de conclusão do curso de graduação						x		1	Caracterização
Curso Grad.	Nome do curso da graduação						x		1	Caracterização
Turma Pref	Turma do aluno no curso da FGV/EAESP					x			1	Caracterização
Ideal conclusão	Ano ideal para a conclusão do curso, com base na data de início					x			1	Caracterização
Série na Grade	Indica qual semestre do curso o aluno está					x			1	Caracterização
Período Letivo	Período letivo em que o aluno está matriculado					x			1	Caracterização
Trancamentos	Número de vezes que o aluno trancou o curso					x			1	Caracterização
Intercambio	Indica se o aluno realizou intercâmbio					x			1	Caracterização
Série	Série do aluno					x		x	2	Caracterização
Série Calculada	Série calculada pelo sistema a partir da data de início do curso					x			1	Caracterização
Data Encerramento	Data de desligamento do aluno						x		1	Caracterização
Crédito por disciplina	Número de créditos que o aluno recebe ao ser aprovado na disciplina							x	1	Caracterização
Notas das disciplinas	Média final da disciplina							x	1	Caracterização
Tipo de Ingresso	Maneira como o aluno ingressou na FGV/EAESP podendo ser transferência, duplo diploma, vestibular, etc				x				1	Caracterização
Paga Inscrição	Indica se o aluno pagou a inscrição para o vestibular		x						1	Caracterização
Isento	Indica se o aluno é isento ao pagamento da inscrição do vestibular		x						1	Caracterização
Questão 01	Você já prestou vestibular para a FGV-EAESP anteriormente? Sim/Não		x						1	Potencial variável
Questão 02	Se sim, chegou a ser convocado para matrícula?		x						1	Não presente em todos
Questão 03	Para que instituições você já prestou ou pretende prestar vertibular neste ano?		x						1	Não presente em todos
Questão 04	Com relação a estudar na FGV-EAESP, você se considera: muito indeciso, indecido, decidido, muito decidido		x						1	Potencial variável
Questão 05	Em qual das faixas abaixo, você estima estar na soma da renda mensal total da sua família?		x						1	Potencial variável
Questão 06	Qual é o nível de instrução de seu pai?		x						1	Potencial variável
Questão 07	Qual é o nível de instrução de sua mãe?		x						1	Potencial variável
Questão 08	Você observou divulgação sobre o vestibular da FGV em algum dos seguintes meios?		x						1	Potencial variável
Nome Disciplina	Nome da disciplina oferecida pela FGV/EAESP no curso de administração de empresas							x	1	Caracterização
Código Disciplina	Código da disciplina oferecida pela FGV/EAESP no curso de administração de empresas							x	1	Caracterização
Crédito Disciplina	Número de créditos da disciplina							x	1	Caracterização
Departamento Disciplina	Departamento da FGV/EAESP responsável por ministrar a disciplina							x	1	Caracterização
Situação do aluno	Situação do aluno em relação à disciplina, podendo ser: aprovado, reprovado por falta, reprovado por nota, trancado							x	1	Variável classificadora
Tipo de disciplina	Indica se a disciplina é obrigatória ou optativa							x	1	Caracterização
Total de informações		49	15	35	24	16	33	15		

Conclusão

APÊNDICE B - Descritivo das variáveis analisadas

Nome da Variável	Descritivo	Qualitativa / Quantitativa	Tipo de variável	Tipo de Variável	Classes iniciais	Categoria Transformada	Classes finais	Nome da variável Dummy	Utilizado nos classificadores?
Bom aluno	Indica 1 se o aluno possui até 1 reprovação por nota e 0 se o aluno possui mais de 1	quantitativa	discreta	classificadora			DP≤1 DP>1	Bom Aluno?	Sim
Prova 1 bruta - Matemática Fase 01	Nota da prova de matemática da fase 01	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 5 inclusive Entre 5 e 6 inclusive Entre 6 e 7 inclusive Entre 7 e 8 inclusive Entre 8 e 9 inclusive Acima de 9	 VP1-D5a6 VP1-D6a7 VP1-D7a8 VP1-D8a9 VP1-DA9	 Sim Sim Sim Sim Sim
Prova 2 bruta - Português Fase 01	Nota da prova de língua portuguesa da fase 01	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 4 inclusive Entre 4 e 5 inclusive Entre 5 e 6 inclusive Entre 6 e 7 inclusive Entre 7 e 8 inclusive Acima de 8	 VP2-D4a5 VP2-D5a6 VP2-D6a7 VP2-D7a8 VP2-DA8	 Sim Sim Sim Sim Sim
Prova 3 bruta - Inglês Fase 01	Nota da prova de língua inglesa da fase 01	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 5 inclusive Entre 5 e 6 inclusive Entre 6 e 7 inclusive Acima de 7	 VP3-D5a6 VP3-D6a7 VP3-DA7	 Sim Sim Sim
Prova 4 bruta - Ciências Humanas Fase 01	Notal da prova de ciências humanas da fase 01	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 4 inclusive Entre 4 e 7 inclusive Entre 7 e 8 inclusive Acima de 8	 VP4-D4a7 VP4-D7a8 VP4-DA8	 Sim Sim Sim
Prova 5 bruta - Matemática Aplicada Fase 02	Nota da prova de matemática aplicada da fase 02	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 5 inclusive Entre 5 e 6 inclusive Entre 6 e 7 inclusive Entre 7 e 8 inclusive Acima de 8	 VP5-D5a6 VP5-D6a7 VP5-D7a8 VP5-DA8	 Sim Sim Sim Sim
Prova 6 bruta - Redação Fase 02	Nota da redação da fase 02	quantitativa	contínua	previsora	Faixas, variando de 1 em 1		Abaixo de 5 inclusive Entre 5 e 6 inclusive Acima de 6	 VP6-D5a6 VP6-DA6	 Sim Sim
Idade	Idade do aluno ao ingressar no curso de Administração de empresas de FGV-EAESP	quantitativa	contínua	previsora	17 a 18 anos inclusive 18 a 19 anos inclusive 19 a 20 anos inclusive acima de 20 anos		Entre 17 e 18 inclusive Entre 18 e 19 inclusive Entre 19 e 20 inclusive Acima de 20	 VIE - D18a19 VIE - D19a20 VIE - DMaior20	 Sim Sim Sim
Região	Região do país onde concluiu o ensino médio, baseado no primeiro dígito do CEP. Caso o aluno cursou o ensino médio no exterior, a sua classificação é "Outro País"	qualitativa	nominal	previsora	Grande SP SP Interior Outras regiões Outro País		Outros Grande SP SP Interior	 VRE - DGSP VRE - DSPI	 Sim Sim
Tipo instituição ensino médio	Tipo de instituição do ensino médio	qualitativa	nominal	previsora	Particular Pública Não definido		Pub_ND Particular	 VET-DPA	 Sim
Questão 01	Você já prestou vestibular para a FGV-EAESP anteriormente?	qualitativa	nominal	previsora	Sim Não		Não Sim	 VQ1	 Sim
Questão 04	Com relação a estudar na FGV-EAESP, você se considera:	qualitativa	ordinal	previsora	Muito indeciso Indeciso Decidido Muito decidido		Mindeciso_Indeciso Decidido Muito decidido	 VQ4-DDE VQ4-DMD	 Sim Sim
Questão 05	Em qual das faixas abaixo, você estima estar na soma da renda mensal total da sua família?	qualitativa	ordinal	provisora	Abaixo de R$ 4.650,00 De R$ 4.651,00 a 9.300,00 De R$ 9.301,00 a 18.600,00 De R$ 18.601,00 a 27.900,00 De R$ 27.901 a 37.200,00 Acima de R$ 37.201,00	Faixa 01 Faixa 02 Faixa 03 Faixa 04 Faixa 05 Faixa 06	Faixa 01_02 Faixa 03 Faixa 04 Faixa 05 Faixa 06	 VQ5-DF3 VQ5-DF4 VQ5-DF5 VQ5-DF6	 Sim Sim Sim Sim
Questão 06	Qual é o nível de instrução de seu pai?	qualitativa	ordinal	previsora	Não freqüentou a escola Ensino fundamental completo (1o grau) Ensino médio completo (2o grau) Graduação completa (ensino superior) Pós-graduação completa (especialização, MBA) Pós-graduação completa (mestrado) Pós-graduação completa (doutorado, PhD)		NF_Grau1_Grau2 Graduação completa (ensino superior) Pós-graduação completa (especialização, MBA) Mestrado_Doutorado	 VQ6-DSP VQ6-DPO VQ6-DMD	 Sim Sim Sim
Questão 07	Qual é o nível de instrução de sua mãe?	qualitativa	ordinal	previsora	Ensino fundamental completo (1o grau) Ensino médio completo (2o grau) Graduação completa (ensino superior) Pós-graduação completa (doutorado, PhD) Pós-graduação completa (especialização, MBA) Pós-graduação completa (mestrado)		Grau1_Grau2 Grau3_Pos Mestrado_Doutorado	 VQ7-DGSP VQ7-DEMD	 Sim Sim
Questão 08	Você observou divulgação sobre o vestibular da FGV em algum dos seguintes meios?	qualitativa	nominal	previsora	QUESTÃO 8 - Internet QUESTÃO 8 - Não vi nenhuma divulgação QUESTÃO 8 - Outros QUESTÃO 8 - Jornal e Revista QUESTÃO 8 - Cartaz, Panfleto e Mala Direta QUESTÃO 8 - Cinema e TV QUESTÃO 8 - Feiras e Eventos QUESTÃO 8 - Cursinho, Colégio, Amigos e Familiares	VQ8-01 VQ8-02 VQ8-03 VQ8-04 VQ8-05 VQ8-06 VQ8-07 VQ8-08	Internet Não vi nenhuma divulgação Outros Jornal e Revista Cartaz, Panfleto e Mala Direta Cinema e TV Feiras e Eventos Cursinho, Colégio, Amigos e Familiares	VQ8-01 VQ8-02 VQ8-03 VQ8-04 VQ8-05 VQ8-06 VQ8-07 VQ8-08	Sim Sim Sim Sim Não Sim Sim Sim
Ensino médio período	Período do dia que o aluno cursou o ensino médio	qualitativa	nominal	previsora	Manhã Tarde Integral		M_I_T M_I_T		Não Não

APÊNDICE C - Análise Bivariada para as variáveis dummies

	Classificação				
Idade	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Entre 17 e 18 inclusive	29%	19%	1,52	0,419	Entre 17 e 18 inclusive
Entre 18 e 19 inclusive	47%	43%	1,08	0,075	Entre 18 e 19 inclusive
Entre 19 e 20 inclusive	17%	23%	0,73	-0,314	Entre 19 e 20 inclusive
Acima de 20	8%	14%	0,52	-0,652	Acima de 20
Total	100%	100%			

	Classificação				
Região	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Grande SP	70%	74%	0,94	-0,062	Grande SP
SP Interior	19%	15%	1,24	0,212	SP Interior
Outros	12%	11%	1,09	0,084	Outros
Total	100%	100%			

	Classificação				
Período	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Manhã	83%	83%	1,01	0,005	M_I_T
I_T	17%	17%	0,98	-0,024	M_I_T
Total	100%	100%			

	Classificação				
Tipo de instituição	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Particular	96%	98%	0,99	-0,015	Particular
Pub_ND	4%	2%	1,63	0,492	Pub_ND
Total	100%	100%			

	Classificação				
Prova 1 bruta - Matemática Fase 01	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 5 inclusive	4%	8%	0,44	-0,810	Abaixo de 5 inclusive
Entre 5 e 6 inclusive	9%	15%	0,61	-0,494	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	10%	16%	0,67	-0,399	Entre 6 e 7 inclusive
Entre 7 e 8 inclusive	29%	30%	0,96	-0,038	Entre 7 e 8 inclusive
Entre 8 e 9 inclusive	20%	15%	1,30	0,264	Entre 8 e 9 inclusive
Acima de 9	29%	16%	1,74	0,554	Acima de 9
Total	100%	100%			

	Classificação				
Prova 2 bruta - Português Fase 01	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 4 inclusive	7%	11%	0,68	-0,389	Abaixo de 4 inclusive
Entre 4 e 5 inclusive	8%	11%	0,74	-0,298	Entre 4 e 5 inclusive
Entre 5 e 6 inclusive	29%	30%	0,94	-0,058	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	18%	17%	1,02	0,016	Entre 6 e 7 inclusive
Entre 7 e 8 inclusive	25%	21%	1,21	0,189	Entre 7 e 8 inclusive
Acima de 8	13%	9%	1,37	0,314	Acima de 8
Total	100%	100%			

	Classificação				
Prova 3 bruta - Inglês Fase 01	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 5 inclusive	7%	13%	0,55	-0,592	Abaixo de 5 inclusive
Entre 5 e 6 inclusive	16%	17%	0,95	-0,049	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	13%	15%	0,87	-0,144	Entre 6 e 7 inclusive
Entre 7 e 8 inclusive	31%	27%	1,13	0,124	Acima de 7
Entre 8 e 9 inclusive	14%	12%	1,18	0,165	Acima de 7
Acima de 9	19%	17%	1,17	0,161	Acima de 7
Total	100%	100%			

Continua

Classificação					
Prova 4 bruta - Ciências Humanas Fase 01	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 4 inclusive	10%	10%	0,97	-0,029	Abaixo de 4 inclusive
Entre 4 e 5 inclusive	9%	11%	0,80	-0,227	Entre 4 e 7 inclusive
Entre 5 e 6 inclusive	25%	31%	0,80	-0,222	Entre 4 e 7 inclusive
Entre 6 e 7 inclusive	14%	17%	0,82	-0,204	Entre 4 e 7 inclusive
Entre 7 e 8 inclusive	30%	24%	1,25	0,225	Entre 7 e 8 inclusive
Acima de 8	13%	8%	1,74	0,553	Acima de 8
Total	100%	100%			

Classificação					
Prova 5 bruta - Matemática Aplicada Fase 02	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 5 inclusive	10%	18%	0,57	-0,558	Abaixo de 5 inclusive
Entre 5 e 6 inclusive	15%	19%	0,79	-0,230	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	24%	26%	0,95	-0,053	Entre 6 e 7 inclusive
Entre 7 e 8 inclusive	24%	22%	1,07	0,069	Entre 7 e 8 inclusive
Acima de 8	26%	15%	1,78	0,577	Acima de 8
Total	100%	100%			

Classificação					
Prova 6 bruta - Redação Fase 02	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Abaixo de 4 inclusive	7%	9%	0,78	-0,254	Abaixo de 5 inclusive
Entre 4 e 5 inclusive	15%	19%	0,81	-0,214	Abaixo de 5 inclusive
Entre 5 e 6 inclusive	28%	29%	0,96	-0,043	Entre 5 e 6 inclusive
Entre 6 e 7 inclusive	32%	28%	1,15	0,135	Acima de 6
Acima de 7	18%	15%	1,19	0,175	Acima de 6
Total	100%	100%			

Classificação					
Questão 01	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	25%	18%	1,38	0,319	Não
Sim	75%	82%	0,92	-0,086	Sim
Total	100%	100%			

Classificação					
Questão 04	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Mindeciso_Indeciso	7%	4%	1,69	0,526	Mindeciso_Indeciso
Decidido	24%	22%	1,09	0,086	Decidido
Muito decidido	68%	73%	0,93	-0,071	Muito decidido
Total	100%	100%			

Classificação					
Questão 05	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Faixa 01_02	22%	19%	1,15	0,139	Faixa 01_02
Faixa 03	23%	23%	1,00	0,004	Faixa 03
Faixa 04	21%	22%	0,98	-0,024	Faixa 04
Faixa 05	12%	16%	0,74	-0,303	Faixa 05
Faixa 06	22%	20%	1,09	0,082	Faixa 06
Total	100%	100%			

Classificação					
Questão 06	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
NF_Grau1	2%	2%	0,87	-0,138	NF_Grau1_Grau2
Ensino médio completo (2o grau)	12%	14%	0,88	-0,124	NF_Grau1_Grau2
Graduação completa (ensino superior)	48%	49%	0,98	-0,019	Graduação completa (ensino superior)
Pós-graduação completa (especialização, MBA)	25%	22%	1,17	0,155	Pós-graduação completa (especialização, MBA)
Mestrado_Doutorado	13%	14%	0,94	-0,058	Mestrado_Doutorado
Total	100%	100%			

Classificação					
Questão 07	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Grau1_Grau2	13%	14%	0,90	-0,102	Grau1_Grau2
Graduação completa (ensino superior)	59%	59%	1,01	0,009	Grau3_Pos
Pós-graduação completa (especialização, MBA)	18%	18%	1,01	0,006	Grau3_Pos
Mestrado_Doutorado	10%	9%	1,08	0,079	Mestrado_Doutorado
Total	100%	100%			

Continua

	Classificação				
Questão 08 - Internet	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	58%	60%	0,96	-0,046	Não
Sim	42%	40%	1,07	0,066	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Não vi nenhuma divulgação	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	86%	84%	1,02	0,024	Não
Sim	14%	16%	0,87	-0,136	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Outros	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	99%	99%	0,99	-0,005	Não
Sim	1%	1%	1,70	0,529	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Jornal e Revista	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	88%	90%	0,98	-0,017	Não
Sim	12%	10%	1,15	0,144	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Cartaz, Planfleto e Mala Direta	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	85%	85%	1,00	0,000	Agrupar
Sim	15%	15%	1,00	-0,001	Agrupar
Total	100%	100%			

	Classificação				
Questão 08 - Cinema e TV	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	97%	96%	1,01	0,008	Não
Sim	3%	4%	0,79	-0,233	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Feiras e Eventos	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	92%	89%	1,03	0,031	Não
Sim	8%	11%	0,75	-0,285	Sim
Total	100%	100%			

	Classificação				
Questão 08 - Cursinho, Colégio, Amigos e Familiares	**DP ≤ 1**	**DP >1**	**Bivariada**	**WOE**	**Novas Classes**
Não	62%	64%	0,98	-0,019	Não
Sim	38%	36%	1,03	0,032	Sim
Total	100%	100%			

Conclusão