Universidade de Brasília
Departamento de Estatística

# MODELAGEM EXPLICATIVA E CLASSIFICATÓRIA POR APRENDIZAGEM SISTÊMICA DO STATUS DISCENTE DO INSTITUTO FEDERAL DE BRASÍLIA

Pedro Carvalho Brom

Apresentada como requisito parcial para
conclusão do curso Bacharelado em Estatística.

Orientador
Prof. Dr. Eduardo Yoshio Nakano

Brasília
2020

Universidade de Brasília
Departamento de Estatística

# MODELAGEM EXPLICATIVA E CLASSIFICATÓRIA POR APRENDIZAGEM SISTÊMICA DO STATUS DISCENTE DO INSTITUTO FEDERAL DE BRASÍLIA

Pedro Carvalho Brom

Apresentada como requisito parcial para
conclusão do curso Bacharelado em Estatística.

Prof. Dr. Eduardo Yoshio Nakano (Orientador)
Universidade de Brasília

Prof. Dr. André Luiz Fernandes Cançado
Universidade de Brasília

Prof. Dr. Guilherme Souza Rodrigues
Universidade de Brasília

Prof. Dr. Jhames Matos Sampaio
Coordenador do Programa de Graduação em Estaística

Brasília, 25 de agosto de 2020

# Resumo

Este trabalho tem como objetivo encontrar variáveis sócio-econômicas que levam o discente a evasão, trancamento ou integralização do curso e se é possível prever essas condições ou o seu êxito antes do início do curso. Os dados se referem aos cursos técnicos do Instituto Federal de Brasília de 2014 a 2017. A modelagem preditiva ocorre dentro da perspectiva estatística com o uso da Regressão Logística e no campo de aprendizagem de máquinas *Random Forest* e Redes Neurais Artificiais, maximizando F1 dentro do critério de *Early Stopping*, consequentemente evitando um sobreajuste.

**Palavras-chave:** Educação, Economia da Educação, Modelos de Classificação.

# Sumário

**1 Introdução** **1**

1.1 Economia educacional . . . . . 3

1.2 Instituto Federal de Brasília e evasão escolar . . . . . 5

**2 Material e Métodos** **7**

2.1 Tarefa de classificação . . . . . 10

2.1.1 Aprendizagem supervisionada . . . . . 10

2.2 Avaliação do modelo preditivo . . . . . 11

2.2.1 Validação cruzada . . . . . 11

2.2.2 Métricas de interesse . . . . . 11

2.3 Aprendizagem Estatística . . . . . 12

2.4 Aprendizagem de Máquinas . . . . . 14

2.4.1 *Random Forest* . . . . . 15

2.4.2 Redes Neurais Artificiais . . . . . 16

**3 Resultados e Discussão** **18**

3.1 Abordagem explicativa . . . . . 20

3.1.1 Prevalência . . . . . 20

3.1.2 Razão de Prevalência . . . . . 21

3.2 Abordagem preditiva . . . . . 21

3.2.1 Regressão Logística e Random Forest . . . . . 22

3.2.2 Redes Neurais Artificiais . . . . . 27

**4 Conclusão** **30**

**Referências** **32**

**Apêndice** **33**

**A : Tabelas** **34**

# Lista de Figuras


1.1 Mapa de calor sobre a rede de informação ligada à evasão escolar. Coloração quente para termos mais frequentes. Fonte: O autor. . . . . 1
1.2 Mapa termos como rede de informação ligada a economia educacional. Fonte: O autor. . . . . 4

2.1 Exemplo de RNA com uma camada oculta. Fonte: O autor. . . . . 17

3.1 Envelopamento para regressão múltipla. Fonte: O autor. . . . . 19
3.2 Construção da modelagem preditiva, primeira etapa. Fonte: O autor. . . . 19
3.3 Construção da modelagem preditiva, segunda etapa. Fonte: O autor. . . . 19
3.4 Validação dos modelos. Fonte: O autor. . . . . 22
3.5 Curva ROC e AUC em treino. Fonte: O autor. . . . . 23
3.6 AUC bootstrap completo. Fonte: O autor. . . . . 25
3.7 AUC bootstrap: Diferença entre modelos. Fonte: O autor. . . . . 25
3.8 Curva ROC e AUC em teste e valid. Fonte: O autor. . . . . 26
3.9 Comparativo dos modelos. Fonte: O autor. . . . . 27
3.10 Matriz de confusão: Comparativo da RNA com RL e RF em valid2. *Output* do comando *caret* :: $confusionMatrix()$. Fonte: O autor. . . . . 28
3.11 ROC e AUC: Comparativo da RNA com RL e RF em valid2. Fonte: O autor. . . . . 29
3.12 Desempenho da RNA: AUC em valid2, comparanto com RL e RF. Fonte: O autor. . . . . 29

# Lista de Tabelas

1.1 Termos mais frequentes em pares. . . . . 2
1.2 Termos mais frequentes em ternas. . . . . 3
1.3 Índice de Gini - Série histórica. . . . . 4

2.1 Matriz de confusão. . . . . 12

3.1 Escores dos modelos em treino, com o ponto de corte $c = 0.5$. . . . . 23
3.2 Escores dos modelos em teste. . . . . 24
3.3 Escores dos modelos em valid. . . . . 24

A.1 Prevalência - parte I. . . . . 35
A.2 Prevalência - parte II. . . . . 36
A.3 Razão de Prevalência. . . . . 37

# 1 Introdução

Seguindo a metodologia Teoria do Enfoque Metaanalítico Consolidado (TEMAC)[1], o levantamento realizado em setembro de 2019 na Web of Science, indicam 163 artigos tratando sobre a temática de evasão escolar. Percebe-se que o assunto ainda é recente na linha de Economia Educacional[1], subárea da Educação. A Figura 1.1, apresenta um mapa de calor sobre a rede de informação ligada à evasão. Esse mapa foi construído por meio dos artigos publicados entre 2015 e 2019 e sugere que ainda não existe uma rede principal conectando os assuntos relacionados.

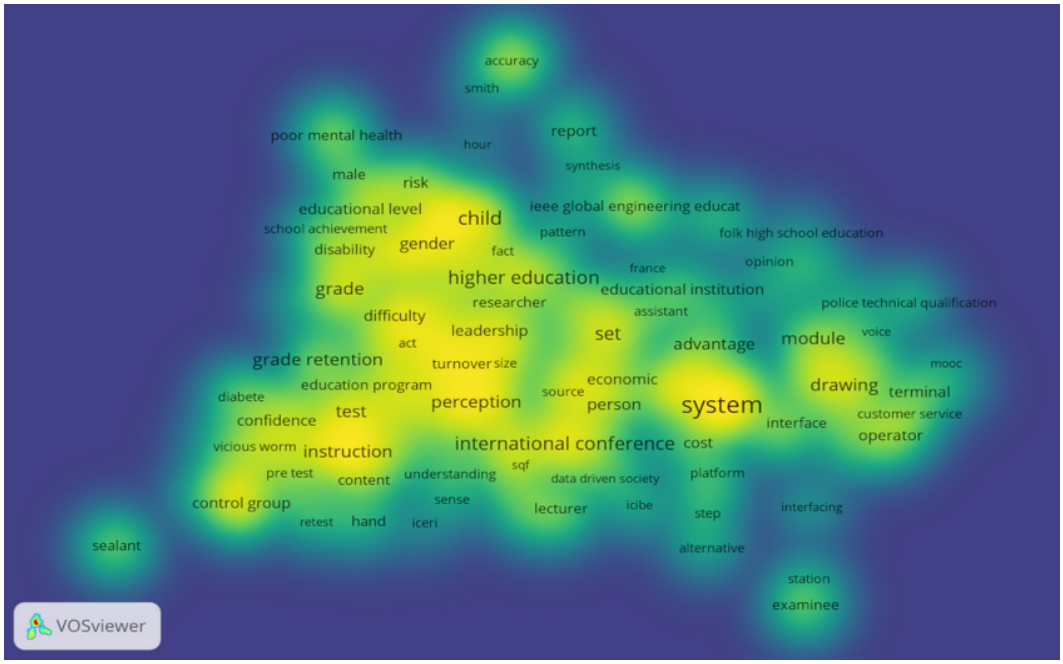


Figura 1.1: Mapa de calor sobre a rede de informação ligada à evasão escolar. Coloração quente para termos mais frequentes. Fonte: O autor.

[1]Código: 70801053. Fonte: https://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-conhecimento-avaliacao.

Uma forma de avaliação sobre o problema da evasão aproveitando os títulos e resumos dos artigos utilizados no TEMAC é técnica *n-gram* para mineração de texto, que utiliza o contexto de termos mais frequentes e a ordem em que ocorrem. Como exemplo temos: *profession qualif*, logo o primeiro termo, *profession* figura nos textos, nesta ordem, junto com *qualif*, formando então um *2-gram*.

Assim, os 10 termos mais frequentes conectados em pares indicam que os estudos atuais seguem uma tendência de termos relevantes como: qualificação profissional, retenção da série, grupos de controle, saúde mental, estudos independentes, desenvolvimento profissional, ganho de instrução e o efeito da retenção, todos no contexto da evasão escolar. É importante ter em mente que a informação atual só reflete o passado e não oferece garantia de tendência futura, portanto serve apenas para entender o contexto dos últimos anos até 2019. Mais detalhes podem ser verificados nas tabelas 1.1 e 1.2. Deste modo, entende-se que a temática evasão/retenção escolar ainda que antiga, de acordo com a literatura mais dos últimos 5 anos, segue neste processo um tanto tímida ao se considerar uma perspectiva quantitativa por leitura estatística. Um ponto notável é a existência estudos longitudinais avaliando desde a condição socioeconômica até saúde mental dos participantes, ou seja existem estudos detalhados sob o olhar das humanas.

| Termos | Tradução livre | Frequência |
|---|---|---|
| *profession qualification* | qualificação profissional | 49 |
| *higher education* | ensino superior | 30 |
| *high school* | colegial | 21 |
| *grade retention* | retenção da série | 19 |
| *education institute* | instituto de educação | 14 |
| *school student* | estudante | 14 |
| *control group* | grupo de controle | 13 |
| *teacher education* | formação de professores | 13 |
| *education program* | programa educacional | 12 |
| *school teacher* | professor da escola | 12 |

Tabela 1.1: Termos mais frequentes em pares.

Ao considerar esses fatores em relação ao resultado, aprovação, preferencialmente dentro do prazo estipulado no Projeto Pedagógico do Curso (PPC), espera-se, ao menos dentro da perspectiva administrativa que o desenlace gerado pelo investimento do governo (instituições públicas) ou o retorno financeiro (instituições particulares) sejam visíveis tanto para os gestores quanto para a comunidade interna/externa. Portanto é razoável entender que existe uma relação natural do capital cultural descrito por Bourdieu[2] proporcionando um melhor posicionamento social ocorrendo em paralelo da eficiência da máquina administrativa. Seguindo esta linha de pensamento entende-se que há uma dinâmica um pouco mais abrangente ocorrendo na perspectiva da economia educacional.

Nestes termos, este trabalho tem como objeto de estudo os fatores que podem explicar a prevalência das condições que levam ao fracasso/sucesso escolar assim como a elaboração de modelos preditivos do status educacional para tomada de decisão seja como medida preventiva ou pesquisa-ação, o que for adequado para a gestão.

| Termos | Tradução livre | Frequência |
|---|---|---|
| *poor mental health* | saúde mental precária | 8 |
| *descript independent claim* | descrição de reivindicação independente | 7 |
| *detail descript independent* | descrição detalhada independente | 7 |
| *develop professional qualification* | desenvolver qualificação profissional | 6 |
| *experiment control group* | grupo de controle do experimento | 6 |
| *nation professional qualification* | qualificação profissional da nação | 6 |
| *novelti method involviment* | envolvimento do método novelti | 6 |
| *toward postgraduate education* | em direção à educação de pós-graduação | 6 |
| *activity enrich instruction* | Instrução de enriquecimento de atividade | 5 |
| *effect grade retention* | efeito do grau de retenção | 5 |

Tabela 1.2: Termos mais frequentes em ternas.

## 1.1 Economia educacional

Colocando a leitura da temática Economia Educacional na linha do tempo (2005-2019), podemos observar que educação, pesquisa e economia são pontos de interesse da América Latina nos anos próximos de 2015. Atualmente, de acordo com figura 1.2, o foco está ligado a economia comportamental, adolescência e jovens adultos. Aspectos de desigualdade andam juntos com a economia regional e educação superior. Essas ponderações podem ser avaliadas diretamente na Figura 1.2 os termos se relacionam como rede de informação ligada a economia educacional.

Cláudio de Moura Castro, autor de Investimento em educação no Brasil: comparação de três estudos disponível no repositório do IPEA[3], detalha aspectos da rentabilidade de investimentos na educação. Em sua primeira concepção faz-se uma crítica da agregação e confiabilidade dos dados, comentando a importância de uma pesquisa de campo bem conduzida para garantir qualidade da informação ou a consulta de dados fidedignos.

Neste estudo o autor pondera a taxa de retorno da educação, configurada no quanto que se ganha por faixa etária. A separação dos dados leva em conta perfis que variam desde analfabetismo, até o universitário completo, indicando que os maiores investimentos eram em 1969 para o primário, seguindo por colegial e a formação universitária. Em termos atuais, são equivalentes ao Ensino Fundamental I e II, Ensino Médio e Superior.

Vale observar que o investimento na educação está relacionado ao conceito de capital humano/intelectual e a capacidade de gerar renda, gerando efeito indireto na desigualdade

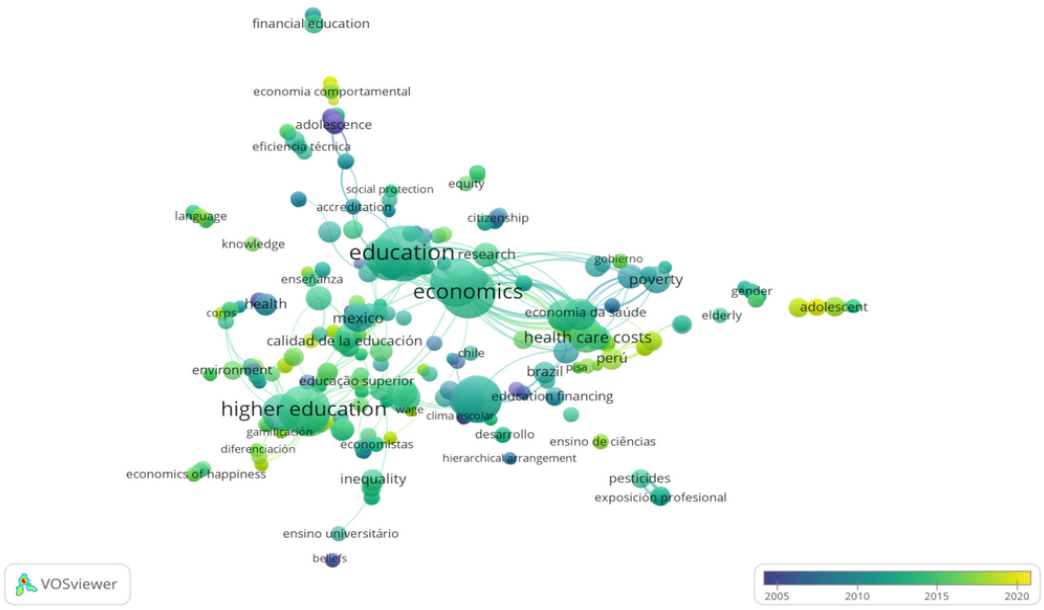


Figura 1.2: Mapa termos como rede de informação ligada a economia educacional. Fonte: O autor.

social[4]. Sem estender muito sobre o assunto, a desigualdade social pode ser medida pelo índice de Gini[2]. Os autores BONELLI e RAMOS[5] informam o índice de Gini nas décadas de 70, 80 e parcialmente de 90 e 00[3] o IBGE[6] complementa a informação mostrando desequilíbrio histórico.[4]

| Década | Índice |
|---|---|
| 60 | 0.500 |
| 70 | 0.568 |
| 80 | 0.590 |
| 90 | 0.574 |
| 00 | 0,539 * |

Tabela 1.3: Índice de Gini - Série histórica.

O ponto em questão se resume com um índice de Gini alto, configurando o Brasil como a Nação da desigualdade. Essas considerações alinhavam o ciclo que retroalimenta

[2]O índice Gini assume valores entre 0 e 1. O valor zero indica completa igualdade (todas as as pessoas têm exatamente a mesma renda) e o valor 1 corresponde a desigualdade completa (uma única pessoa recebe toda a renda e os demais não têm renda alguma).

[3]Valor médio de cada década.

[4]Na tabela: Índice de Gini - Série histórica. a notação * indica o valor médio da década. A contagem de tempo está limitada até 2007.

a necessidade de formação profissional com o arranjo produtivo local. Neste cenário as Escolas Politécnicas fundadas em 1810 inicialmente objetivando a formação de oficiais de artilharia e, nos dias atuais como Institutos Federais, estão com uma nova missão de *Oferecer ensino, pesquisa e extensão no âmbito da Educação Profissional e Tecnológica, por meio da inovação, produção e difusão de conhecimentos [...]*[7]. Deste modo segue uma leitura do papel do Instituto Federal de Brasília e a evasão escolar.

## 1.2 Instituto Federal de Brasília e evasão escolar

O Instituto Federal de Brasília (IFB) foi criado pela lei n o 11.892, de 29 de dezembro de 2008. O IFB é composto por uma Reitoria e 10 campi distribuídos pelo Distrito Federal: Brasília, Ceilândia, Estrutural, Gama, Planaltina, Riacho Fundo, Samambaia, São Sebastião, Taguatinga e Recanto das Emas. Conta com cinco Pró-Reitorias: de Administração (PRAD), de Ensino (PREN), de Extensão e Cultura (PREX), Gestão de Pessoas (PRGP) e de Pesquisa e Inovação (PRPI), atuando no Ensino Básico, Técnico e Tecnológico.

Ainda que o IFB tenha toda a estrutura para um bom funcionamento, encontramos o problema da evasão escolar[5] presente em grande parte das instituições de ensino do Brasil. No Distrito Federal, cerca de 23% dos estudantes do IFB se encontram no que se define como grupo de Risco Escolar, composto por: Cancelado, Evadido, Trancado ou Integralizado[6]. Esta taxa de evasão pode ter relação direta com a condição socioeconômica discente.

Em 2014 foi implementado o Sistema de Gerenciamento de Processo Seletivo (SGPS), que permite a coleta de informações no ato de inscrição para a vaga pretendida. Neste questionário são propostas questões sobre o núcleo familiar, trabalho, profissão, endereço, uso de tecnologias/internet e até mesmo afinidade cultural, por exemplo. O sistema está em constante atualização e nos últimos anos o mesmo passou de 29 perguntas para cerca de 40 sendo todas de múltipla escolha. No mesmo ano foi implementado o Sistema de Gestão Acadêmico (SGA), que permite a busca do status discente separado em dois grupos recategorizados: 1. Risco Escolar (23%): Cancelado, Evadido, Trancado ou Integralizado e, 2. Êxito Escolar (77%) Matriculado ou Formado[7]. Deste modo podemos cruzar informações do SGPS e SGA para tentar prever padrões de status discente antes da matrícula. Em relação à forma de ingresso, para os cursos de nível superior utilizam-se o Sistema

[5]http://portal.inep.gov.br/artigo/-/asset$_p$*ublisher*/*B*4*AQV*9*zFY*7*Bv*/*content*/*inep* − *divulga* − *dados* − *ineditos* − *sobre* − *fluxo* − *escolar* − *na* − *educacao* − *basica*/21206.

[6]Dados do autor.

[7]Neste trabalho considera-se matriculado como de sucesso e a perspectiva é que em trabalhos futuros considere-se apenas formado.

de Seleção Unificada (SiSU), já para os cursos técnicos subsequentes e técnicos integrados (Ensino Médio ou PROEJA[8]) a forma ingresso segue por Amostragem Estratificada (sorteio), devido às cotas dos grupos de vulnerabilidade social.

Acredita-se que nestes termos é possível avaliar a condição do estudante antes mesmo de se iniciarem os estudos, pois é natural reconhecer a possibilidade de padrão oculto que pode ser aprendido na perspectiva estatística ou de máquinas gerando modelos preferencialmente interpretáveis e com boa capacidade de predição. Por exemplo, o artigo *Big Data in Educational Administration: An Application for Predicting School Dropout Risk*, SORENSEN (2018)[8] descreve o uso de aprendizagem de máquinas para avaliar o risco da evasão escolar usando dados longitudinais. Ao todo utilizou-se 74 variáveis que vão desde indicadores comportamentais a características socioeconômicas e demográficas. Como resultado o modelo preditivo tem a acurácia de 90%. O ponto alto do estudo é o esclarecimento das técnicas utilizadas para predição funcionando disjunto da ação política, ou seja, o artigo apenas indica o potencial de novas formas para reunir informações construindo o conhecimento sobre a temática evasão escolar.

[8]Programa Nacional de Integração da Educação Profissional com a Educação Básica na Modalidade de Educação de Jovens e Adultos (PROEJA).

# 2 Material e Métodos

Os dados do Sistema de Gerenciamento Acadêmico (SGA) e Sistema de Gerenciamento de Processo Seletivo (SGPS) foram solicitados em dezembro de 2016 seguindo os trâmites administrativos e finalizando o processo de acesso em abril de 2017. Foram separados em dois arquivos, um para cada sistema, na forma de *dump* do MySQL. Por se tratar da primeira solicitação, houve uma certa demora para liberação, pois foi necessário consultar o jurídico para o acesso das bases com a finalidade de pesquisa. Vale observar que o pedido foi anterior à lei nº 13.709, de 14 de agosto de 2018, mais conhecida como Lei Geral de Proteção de Dados Pessoais (LGPD) que dá providências em seu Art. 2º sobre *A disciplina da proteção de dados pessoais tem como fundamentos: I - o respeito à privacidade; [...] V - o desenvolvimento econômico e tecnológico e a inovação [...]* e no Art. 7º, que regula a condição de uso para fins acadêmicos: *O tratamento de dados pessoais somente poderá ser realizado nas seguintes hipóteses: [...] IV [...] a anonimização dos dados pessoais [...].*

Para tanto os dados recebidos foram pré-tratados, cruzando as tabelas de *Standard Query Language* (SQL) e garantindo o anonimato somente após a publicação da lei nº 13.709. A base de dados do SGA tem um total de 26255 registros de matrícula únicos e 25 variáveis, já a do SGPS totalizam 180955 registros de CPF, sendo que existem repetições de processos seletivos anteriores, portanto são 179659 registros únicos. Em relação a quantidade de variáveis, temos um total de 43. Foram eliminadas observações com algum tipo de inconsistência, por exemplo: indivíduo diz que não está trabalhando, porém em outro item do questionário do SGPS informa que o seu turno de trabalho é pela tarde. Outro tipo de modificação foi em relação aos eventos raros, como uma lista de necessidades específicas que isoladamente não chegavam a 5 observações, assim, considerou-se apenas duas categorias Alguma necessidade e Nenhuma necessidade. Observações relacionadas aos cursos técnicos integrados ao ensino médio ou de jovens e adultos foram removidos. Cursos identificados por Edificações, Comércio, Eventos, Eletromecânica, Alimentação, Automóveis, Eletrônica e Panificação foram agrupados por terem poucas observações na categoria Cursos com poucos alunos.

Após o cruzamento das duas bases obtivemos um total de 4432 registros únicos e completos, tratando-se de um *subset* apenas dos cursos técnicos subsequentes, pois estamos

interessados na propriedade de aleatorização conforme detalhado na introdução, tópico Instituto Federal de Brasília e evasão escolar. Ao todo são 27 questões compatíveis com as alterações ocorridas desde 2014 reorganizadas de tal modo que as categorias sejam mutuamente excludentes. Essas questões são descritas a seguir (descrição/ categorias ou intervalo/ **nome da variável**):

1. Condição final do estudante./ Categórica nominal: Êxito; Sem Êxito./ **status**.
2. Sexo do respondente./ Categórica nominal: Feminino; Masculino./ **sexo**.
3. Grau de instrução./ Categórica ordinal: Ensino Médio Completo; Superior Incompleto; Superior Completo ou Pós Graduação./ **grau_instr**.
4. Raça./ Categórica nominal: Branca; Indígena ou Amarela; Parda; Preta e NI[1]./ **raca**.
5. Estado civil./ Categórica nominal: Separado(a) ou Divorciado(a); Solteiro(a); União Estável (UE) ou Casado(a); Viúvo(a); NI./ **est_civil**.
6. Curso./ Categórica nominal: Administração; Agro+[2]; Controle Ambiental; Cozinha; Equipamentos Biomédicos; Informática; Infraestrutura Escolar; Logística; Meio Ambiente; Móveis; Multimeios Didáticos; Química; Secretariado; Segurança; Serviços e Vestuário; Outros cursos com poucos alunos./ **curso**.
7. Unidade educacional./ Categórica nominal: *campi* Brasília (CBRA), Ceilândia (CCEI), Estrutural (CEST), Gama (CGAM), Planaltina (CPLA), Riacho Fundo (CRFI), Samambaia (CSAM), São Sebastião (CSSE), Taguatinga (CTAG), Taguatinga Centro (CTGC)[3]./ **campus**.
8. Forma do curso./ Categórica nominal: Componente ou Outros; Modular./ **forma_curso**.
9. Modalidade./ Categórica nominal: À distância; Presencial; Semi-Presencial./ **modal**.
10. Área do curso./ Categórica nominal: Ciência Exatas e da Terra ou Engenharias; Ciências Agrárias; Ciências da Saúde Ciências Sociais; Aplicadas ou Ciência Humanas./ **area**.

[1]NI significa Não Informado pelo respondente no SGPS ou não informado no SGA.
[2]Todos os cursos relacionados ao Agronegócio.
[3]Optou-se pelo uso desta nomenclatura, mas o referido *campus* foi desativado para criação do *campus* Recanto das Emas

11. Eixo do curso./ Categórica nominal: Ambiente; Saúde e Segurança; Apoio Educacional; Controle e Processos Industriais; Desenvolvimento Educacional e Social; Gestão e Negócios; Hospitalidade e Lazer; Informação e Comunicação; Infra-estrutura; Produção Alimentícia; Produção Industrial; Recursos Naturais./ **eixo**.

12. Origem escolar do estudante./ Categórica nominal: Particular; Pública; NI./ **ori_escol**.

13. Necessidade específica./ Categórica nominal: Alguma necessidade; Nenhuma necessidade./ **nec_esp**.

14. Carga horária do curso./ Categórica ordinal: Menor do que 500 horas-relógio]; Maior ou igual a 500 horas-relógio; NI./ **ch_curso**.

15. Idade em anos completos./ Numérica discreta: Valores: mínimo 18 e máximo 83./ **anos_comp**.

16. Motivo da escolha do curso./ Categórica nominal: Realização profissional ou Profissionalização; Relação com Trabalho ou Complementação; Sem relação Profissional./ **mot_escolha**.

17. Influência da escolha./ Categórica nominal: Gratuidade ou Conveniência; Qualidade do curso; NI./ **influ_escolha**.

18. O respondente tem computador em casa./ Categórica nominal: Não; Sim; NI./ **pc_casa**.

19. O respondente tem internet em casa./ Categórica nominal: Não; Sim; NI./ **internet_casa**.

20. Qual o principal meio de comunicação?/ Categórica nominal: Internet; Outras Mídias; TV; NI./ **meio_atual**.

21. Ocupação do respondente./ Categórica nominal: Detém alguma renda sem trabalho; Sem trabalho; Setor Privado com ou sem Vinculo ou Servidor Publico; NI./ **ocup**.

22. Trabalha na área do curso?/ Categórica nominal: Não; Sim; NI./ **trab_area_curso**.

23. Turno do trabalho./ Categórica nominal: Manhã; Manhã e Tarde; Noite; Outro; Sem Trabalho; Tarde; NI./ **turno_trab**.

24. Principal meio de transporte./ Categórica nominal: Outro; Particular; Publico; NI./ **transporte**.

25. Definida por quantidade de filhos./ Categórica ordinal: 0; 1; 2; 3+; NI./ **qtd_filhos**.

26. Definida pelo total de moradores na mesma moradia./ Categórica ordinal: 1; 2; 3; 4; 5; 6+; NI./ **qtd_moradores**.

27. Definida pelo número de salários mínimos da família[4]./ Categórica ordinal: Até 1; Mais de 1 e menos que 2; Mais que 2 e menos do que 5; 5 ou +; NI./ **renda_fam**.

O desenvolvimento do trabalho está pautado em duas abordagens:

1. Aprendizagem Estatística.
2. Aprendizagem de Máquinas.

A finalidade da aprendizagem estatística está no reconhecimento de padrões associados à prevalência escolar quanto ao status escolar. Já para fins preditivos serão construídos modelos lineares generalizados que serão avaliados e validados para serem comparados com os modelos de aprendizagem de máquinas.

## 2.1 Tarefa de classificação

Classificação é o problema de agrupamento por características de um conjunto de treinamento em que, quando obtemos uma nova observação, podemos indicar o grupo mais provável devido ao padrão percebido. Este padrão pode ser aprendido de modo supervisionado quando conhecemos previamente os grupos existentes ou de modo não supervisionado criando a regra de agrupamento sem o conhecimento prévio destas observações. Algoritmo de classificação segue uma aproximação obtida por uma função de reconhecimento da amostra dado um conjunto de características de $n$ elementos. Alguns exemplos de aplicações são: Risco de crédito[9], reconhecimento facial/vocal[10], marketing[11][12], biometria[13], assinatura cognitiva[14], diagnóstico médico antecipado[14] entre outros. Este trabalho está focado apenas em técnicas supervisionadas.

### 2.1.1 Aprendizagem supervisionada

Regras de classificação são desenvolvidas por aprendizagem a partir de amostras, sendo medidas características em relação às diferenças apresentadas[15]. Deste modo criamos uma função que permite avaliar estas informações minimizando o erro de agrupamento[16]. Basicamente temos dois tipos de erro, sendo I conhecido por falso positivo e que ocorre ao indicar que a observação pertence ao grupo de interesse enquanto pertence a outro. O tipo II é o falso negativo decorrente da classificação em outro grupo quando na verdade pertence ao grupo de interesse, podendo um ser mais grave do que o tipo I dependendo da área de aplicação.

[4]As diferentes versões dos questionários traziam intervalos distintos e em alguns casos numérica. Para ajuste desta variável adotou-se o tabelamento do governo vigente de cada ano e reclassificadas corretamente.

Ao classificar, para cada vetor de informações $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ip}), i = 1, 2, ..., n$ observado, avaliamos se pertence a uma classe de interesse o valor retornado de $\hat{Y}$. Por exemplo suponha $Y$ composto por dois grupos $G_1$ se $Y = 1$ (classe positiva) e $G_2$ se $Y = 0$ (classe negativa). No caso da nova observação temos:

$$\begin{cases} \hat{Y}_i \geq c \rightarrow \mathbf{x}_i \in G_1 \\ \hat{Y}_i < c \rightarrow \mathbf{x}_i \in G_2 \end{cases}, c \in [0, 1]; i = 1, ..., n.$$

Em geral é usual $c = 0.5$[5] e pode ser alterado para melhorar a taxa de acerto global. Note que em problemas aplicados teremos $c \in (0, 1)$, pois no caso de $c = 0$ ou $c = 1$ um modelo não é mais necessário. Outro ponto é a função de ativação. Esta retornará uma Função de Distribuição Acumulada (FDA) tal que escolheremos o ponto de corte adequado para separar as categorias.

## 2.2 Avaliação do modelo preditivo

### 2.2.1 Validação cruzada

*Cross Validation* (CV) ou Validação Cruzada é uma estimativa do desempenho geral do modelo treinado em $k$ blocos de informação (partições) escolhidos aleatoriamente da base de dados. Geralmente é uma estimativa ligeiramente menor do desempenho de um modelo treinado em $n$ observações. Este procedimento é realizado treinando o modelo em $k - 1$ blocos e aplicando alguma métrica de interesse no bloco de teste. É feito um rodízio até serem feitos $k$ testes. Caso a média de erros seja pequena então o modelo se mostra bem comportado para a base de dados. Por outro lado se ocorre valores altos então o modelo se mostra sensível a perturbações na base de dados indicando que devemos buscar algum modelo mais robusto. Neste trabalho a base de dados está particionada em treino, teste e validação com 2303, 1152 e 1150 observações respectivamente.

### 2.2.2 Métricas de interesse

Matriz de confusão é uma forma de organizar com uma tabela cruzada o que é predito confrontando com a informação real. Basicamente informa o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Isso permite calcular algumas métricas de interesse para o problema proposto. Por exemplo, na tabela 2.1 temos:

[5]Considere $c \in (0, 1)$ valor fixo, arbitrado pelo pesquisador.

| | | Valores reais | |
|---|---|---|---|
| | | Positivo | Negativo |
| Predição do modelo | Positivo | Verdadeiro Positivo ($VP$) | Falso Positivo ($FP$) |
| | Negativo | Falso Negativo ($FN$) | Verdadeiro Negativo ($VN$) |

Tabela 2.1: Matriz de confusão.

Valor Preditivo Positivo ($VPP$) é a taxa de acerto para o grupo positivo, $VPP = VP/(VP + FP)$. Valor Preditivo Negativo ($VPN$) é a taxa de acerto para o grupo negativo, $VPN = FN/(FN + VN)$. Acurácia ($ACC$) é a taxa global de acertos, $ACC = (VP + VN)/(VP + FP + FN + VN)$.

Sensibilidade ($SENS$) mede a proporção de positivos que estão corretamente identificados, $SENS = VP/(VP + FN)$. Especificidade ($ESPEC$), semelhante à sensibilidade, mas para a proporção de negativos, $ESPEC = VN/(VN + FP)$.

$F1$, que pode ser interpretado como uma média ponderada entre $VPP$ e $SENS$, com valores entre 0 e 1, sendo desejável valores altos. Normalmente a $F1$ é recomendada para dados desbalanceados[6] e seu cálculo é dado por: $F1 = 2 \times VPP \times SENS/(VPP + SENS)$

Veja que algumas dessas informações podem ser organizadas de modo equilibrado através da *Receiver Operating Characteristic* ($ROC$)[7]. A $ROC$ é construída pelo gráfico balanceado entre $VP$, também conhecida como sensibilidade ou probabilidade de identificação, e $FP$, calculada por $1 - ESPEC$. Nestes termos existe um ponto de corte $c \in (0, 1)$ tal que maximize $SENS$ e $ESPEC$, isto é, maximiza *Area Under the ROC Curve* ($AUC$). Os modelos preditivos propostos neste estudo seguem $AUC$ para fins de comparação.

## 2.3 Aprendizagem Estatística

Considerando o cenário de evasão escolar então podemos conceber algumas possibilidades de modelagem explicativa e preditiva por aprendizagem estatística. Regressão Logística, Binomial Negativo e Poisson com Variância Robusta são candidatos naturais. Segundo AGRESTI[17], *As variáveis categóricas são difundidas nas ciências sociais para medir atitudes e opiniões, com categorias como (concordo, discordo), (sim, não).* No caso o estudo considerando que existe o Risco Escolar (1) e Êxito Escolar (0). Aqui entende-se Risco Escolar não no sentido estatístico de Risco Absoluto ou Risco Relativo e sim na perspectiva da administração[8].

[6] Methods for Dealing with Imbalanced Data. https://www.kaggle.com/tboyle10/methods-for-dealing-with-imbalanced-data

[7] Descrito no tópico 2.2.

[8] Entende-se risco em administração como a ocorrência de um evento de modo aleatório e seus impactos subsequentes.

Modelos Lineares Generalizados (MLG) pertencem a uma classe ampla de modelos de regressão que trabalham com variáveis de resposta contínua, discreta ou categórica. Os modelos lineares generalizados basicamente têm três componentes: Aleatório que representa a variável $Y$ com distribuição de probabilidade predefinida e o preditor linear que são as variáveis explicativas intermediadas pela função de ligação que será o valor esperado de $Y$ condicionado ao preditor linear.

O componente aleatório neste problema está ligado ao sucesso ou fracasso do evento status, assim podemos considerar a contagem destes sucessos por distribuição Binomial, Binomial Negativa ou até mesmo por Poisson com os devidos ajustes. Dependendo do critério de amostragem adotado e/ou das propriedades de média, variância e qualidade da aderência/ajustamento definimos a melhor forma de representação dos dados por um modelo probabilístico.

De modo genérico descrevemos $Y \sim D(\mu, \sigma^2)$, com $D$ uma distribuição de probabilidade adequada ao problema. O preditor linear especifica as variáveis que explicam (podem gerar efeito) na função de ligação, $g(\mu(\mathbf{X}))$, com $\mu(\mathbf{X}) = E(Y)$. Sua forma é dada por:

$$g(\mu(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p.$$

A função de ligação depende do problema que está sendo apresentado, por exemplo, quando o componente aleatório segue uma Poisson, a $g(\mu(\mathbf{X}))$ pode ser representada por $log(\mu(\mathbf{X}))$ ou em uma Binomial-Negativa por $log(\mu(\mathbf{X}))$ ou $\sqrt{\mu(\mathbf{X})}$. Neste estudo seguiremos por uma abordagem mista, pois estamos interessados nos efeitos das variáveis assim como avaliar a capacidade de predição.

Em termos técnicos, dizemos que a distribuição da variável aleatória $Y$ pertence à família exponencial unidimensional de distribuições se pudermos escrever sua função de probabilidade (f.p.) ou função de densidade de probabilidade (f.d.p.) como: $f(y|\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\}$, com $b$, $c$ funções reais do parâmetro $\theta$ e $a$, $d$ são funções de $y$. A vantagem da escrita na forma exponencial se dá aos resultados fechados para média e variância:

$$E(Y) = -\frac{c'(\theta)}{b'(\theta)} \quad e \quad Var(Y) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

Note que este trabalho tem como objetivo estudar os efeitos das covariáveis quanto ao status, usando dados de contagem. Para tanto serão abordados os modelos: (i) Poisson estimado por Máxima Verossimilhança (MV) e por Quasi-Verossimilhança (QV) e (ii) Binomial-Negativo. Não será utilizada a Regressão Logística para fins explicativos pois é mais interessante para a administração entender a razão de prevalência do que razão de chances. Além disto justifica-se esta escolha devido aos intervalos de confiança mais

estreitos e sua simplicidade na interpretação.

Em (i), seja o componente aleatório $Y \sim Poisson(\mu)$ e, por MV, $E(Y) = Var(Y) = \mu$. Sendo estimado por QV, $E(Y|X) = e^{X\beta}, Var(Y|X) = \phi e^{X\beta}$, com $\phi = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} r_{pad_{ij}}^2}{n-p}$, com $r_{pad_{ij}}$, resíduo padronizado de Pearson na posição $ij$. Interpretamos $\phi \in (0,1)$ como subdisperso, $\phi = 1$ equidisperso e $\phi > 1$ superdisperso. Em (ii), seja o componente aleatório $Y \sim BN\left(\frac{1}{D}, \frac{D\mu(\mathbf{X})}{D\mu(\mathbf{X})+1}\right)$ tal que $E(Y) = \mu(\mathbf{X})$ e $Var(Y) = \mu(\mathbf{X}) + D\mu(\mathbf{X})^2$, com $D$ parâmetro de dispersão e $\mu(\mathbf{X}), D > 0$.

Para todos os casos considere a função de ligação: $log(\mu(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$, portanto: $\mu(\mathbf{X}) = exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\} = e^{\beta_0} \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times ... \times e^{\beta_p X_p}$. os resultados $e^{\beta_j X_j}, j = 1, 2, ..., p$ refletem o $j$-ésimo efeito sobre o valor esperado para cada unidade de $X_j$. Modelos com distribuição Binomial-Negativa ou modelos Quasi-Poisson servem para casos de superdispersão, ou seja, variância maior do que a de Poisson.

Vale observar que $\mu$ pode representar a proporção de aprovados ou reprovados, por construção. Além disso modelos de Poisson com variância robusta e o Binomial-Negativo têm entrado em uso como alternativa em dados de contagem para o cálculo da Razão de Prevalência[18] e obter $e^{\beta_j X_j}, j = 1, 2, ..., p$ encontramos uma aproximação com viés, porém mais precisa, da Razão de Prevalência para o nível de referência.

Todas essas considerações continuam valendo para fins preditivos e aqui finalizamos com o uso da Regressão Logística e sua respectiva função de ligação, $log\left(\frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$, $\mu(\mathbf{X}) \in (0,1)$ em que $\frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})}$ representa a Razão de Chances do evento de interesse.

## 2.4 Aprendizagem de Máquinas

Aprendizagem de máquinas é semelhante ao raciocínio humano. Em termos simplificados quanto mais lemos e estudamos sobre um assunto, mais fácil é a tarefa de sobre o tema. Como exemplo indica-se o estudo sistêmico da matemática. Repare que com o tempo conseguimos entender como os temas se comunicam e os assuntos estão colecionados de modo intuitivo, pois dentro de nossa cabeça não existe uma regra clara para separação[19].

É certo que podemos estabelecer alguma heurística para sistematizar os assuntos o que facilita a tarefa de classificação, mas sempre aparecerá algo que não será classificado na categoria adequada, pois existe uma certa complexidade que não é linear. Note que cada pessoa terá uma forma de classificar, que naturalmente pode deter alguma similaridade com a outra pessoa, mas sempre existe algo que as diferenciará e não pode ser explicado (caixa preta).

De modo semelhante a máquina aprende algum critério de classificação ou de regressão por aproximações, partindo de algum algoritmo pré-determinado, ou seja, quanto

mais dados são processados, mais fácil será a aproximação da resposta correta, mas não conseguimos explicar os motivos que levam ao resultado de classificação (caixa preta).

Assim como devemos estudar um livro e não decorar, também podemos alinhar o treinamento do modelo de tal forma que ao aparecerem novos exercícios conseguiremos relacionar a informação passada com a atual, podendo acertar ou errar a questão. Se o processo de treinamento for diligente e os dados forem de qualidade então conseguiremos uma taxa de acerto balanceada entre $VP$ e $FP$, isto é, a $AUC$ terá boa cobertura.

Um ponto interessante a se considerar são os modelos disponíveis para treinamento. Naturalmente existem algoritmos que diferem do aprendizado estatístico, por exemplo *Random Forest* (RF) ou uma Rede Neural Artificial (RNA), mas de modo geral os mesmos modelos acabam figurando no meio de aprendizagem de máquinas. O que muda é a abordagem da métrica de treinamento e os facilitadores. Em aprendizagem de máquinas podemos treinar considerando *Area Under Curve* (AUC) com mais facilidade[9] do que na aprendizagem estatística. Em contraparte aprendizagem de máquinas não permite interpretabilidade e o esforço pode ser computacionalmente intensivo se comparado a aprendizagem estatística e nem sempre os resultados são melhores.

### 2.4.1 *Random Forest*

*Random Forest*, desenvolvido por Breiman[20] consiste na modificação da técnica *bootstrap aggregation* em que são calculadas as médias de árvores decisórias não correlacionadas. Cada árvore é construída partindo de um vetor aleatório amostrado de modo independente e identicamente distribuído (i.i.d.). O erro dessa é diminuído quando aumentamos o número de árvores amostradas. Random Forest podem capturar interações mais complexas dos dados e obtendo uma quantidade considerável de árvores decisórias, o viés é relativamente pequeno[21].

Um ponto interessante sobre a construção das árvores é uma consequência por serem identicamente distribuídas (i.d.). Considere o cenário em que o processo é i.d. então a média destas variáveis aleatórias denotada por $B$ são idênticas. Por outro lado a variância do processo será da forma $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$, sendo: correlação positiva aos pares $\rho$, variância desta média $\sigma$ e média $B$. Deste modo se $B$ aumenta então sobra $\rho\sigma^2$, o que no pior cenário se $\rho = 1 \implies \sigma^2$. Por outro lado se as amostras forem i.i.d., teremos $\frac{\sigma^2}{B}$, ou seja, é um método que traz algum ganho livre da composição da amostra. Deste modo i.i.d. o algoritmo diminui a correlação entre as árvores decisórias sem afetar tanto a variância.

O algoritmo *Random Forest* é construído da seguinte forma:

1. Para $b = 1$ até $B$ faça:

[9] Pacote *caret* do R.

(a) Realize uma amostra *bootstrap* $\mathbf{Z}$ de tamanho $N$ da base de treinamento.

(b) Construa uma floresta aleatória $T_b$ da amostra repetindo os seguintes passos para cada nó da árvore até que seja encontrado o nó minimal de tamanho $n_{min}$:

i. Selecione $m < p$ variáveis dentro das $p$ disponíveis.

ii Pegue a melhor variável dentre as $m$ disponíveis para o nó de separação.

iii. Com este nó crie duas ramificações.

2. Crie a saída desta construção de árvores $\{T_b\}_1^B$, com 1: grupo de interesse.

Para realizar uma nova predição de um ponto $x$, dentro da tarefa de classificação, defina por $\hat{Y}_b(x)$ a classe de predição da b-ésima árvore aleatória. Então $\hat{Y}_{RF}^B(x)$ será igual a maioria das proporções de $\{\hat{Y}_b(x)\}_1^B$.

### 2.4.2 Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) a partir das entradas cria combinações lineares (neurônios) e avalia a condição de saída por alguma função de ativação, no caso da tarefa de classificação. Isso gera um método de aprendizado extremamente versátil para problemas em que a separação não é linear.

No problema de aprendizagem supervisionada, de modo genérico dizemos um vetor de entrada $X$ com $p$ componentes e um vetor saída $Y$ como meta da aprendizagem. Considere $\omega_m, m = 1, 2, ..., p$, pesos para os p-vetores do sistema de entrada. Deste modo o sistema pode ser descrito por:

$$f(X) = \sum_{m=1}^{p} g_m(\omega_m^T X),$$

também chamado de modelo aditivo, com função $g_m$ não especificada que direciona o ajuste do vetor $\omega_m$. Essa função $g_m(.)$ pode ser uma curva logística, por exemplo. A escolha dos pesos iniciais pode ser de forma pseudo-aleatória seguindo alguma distribuição de probabilidade conveniente. Então, seguindo as sugestões deste exercício, ocorre $\omega_m^T X$ e aplicamos em $g_m(.)$. O resultado operamos multiplicação nos pesos $\omega^*$, gerando um novo total que deve ser aplicado novamente na função de ativação escolhida, que chamaremos de $\hat{Y}$.

Neste ponto devemos considerar alguma função de perda para comparar com os resultados de $Y$ e realizamos a atualização dos pesos por $\omega_* = X^T\delta$, $\delta = erro.g(\omega_{antigo}^T X)$ e renovamos por $\omega_{novo} = \omega_{antigo}.M + \omega_*.\lambda$, com $M$ definido por momento[10] e $\lambda$ a taxa

[10]É uma taxa preventiva do algoritmo parar em mínimo local.

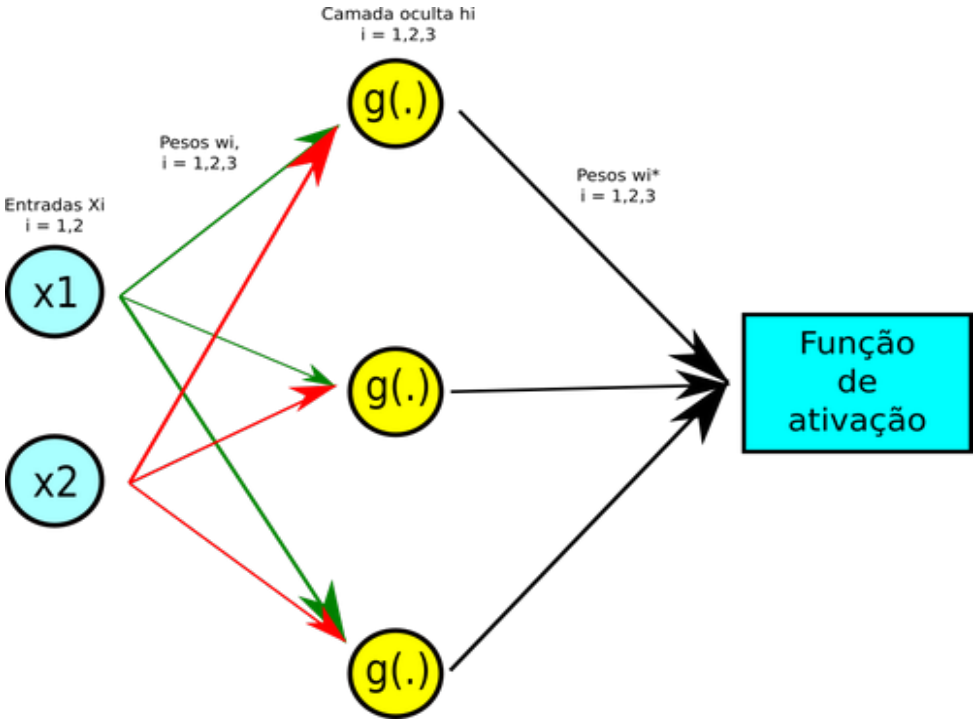


Figura 2.1: Exemplo de RNA com uma camada oculta. Fonte: O autor.

de aprendizagem[11]. Essa atualização deve ser feita tanto os pesos da camada de entrada $\omega_i, i = 1, 2, ..., p$ quanto para a acamada oculta $\omega_i^*, i = 1, 2, ..., p$.

Seu algoritmo é construído da seguinte forma:

1. Inicializamos os vetores de pesos para cada camada com valores pseudo-aleatórios.
2. Baseado nos dados, operamos as funções de ativação e calculamos os erros.
3. Atualizamos os pesos (backpropagation).
4. Pararamos quando o erro for suficientemente pequeno (convergência).

[11]Indica a velocidade em que a RNA converge. Valores de $\lambda$ menores implicam em melhor aprendizagem, porém deixam o algoritmo lento.

# 3 Resultados e Discussão

Os resultados explicativos seguem interpretando a prevalência e razão de prevalência do status de Risco Escolar. A definição de Risco Escolar e significado seguem conforme discutido no capítulo anterior. Para fins de identificação da variável *status* nos resultados, com as categorias sem_exito e exito representam estudantes em Risco Escolar e Êxito Escolar respectivamente. O modelo de Poisson com variância robusta[1] foi escolhido em detrimento do Binomial Negativo para que não tivéssemos problemas de estimação. Portanto, ainda que os resultados possam apresentar pequeno viés, seguem confiáveis em sua interpretação e inferência[2].

Partindo para avaliação do efeito aditivo das variáveis explicativas, foram construídos os modelos de Regressão Logística, Quasi-Poisson e Binomial Negativo. Nestes avaliamos por Monte Carlo o envelopamento do intervalo de confiança a 95%, porém os modelos não se mostraram adequados para fins de inferência devido a baixa qualidade de ajuste. Para a regressão logística tivemos um modelo superajustado deixando 0.35% dos valores simulados fora, sendo que o esperado era de 5%. Nos modelos de Quasi-Poisson e Binomial Negativo, obtivemos 83.78% e 99.13% dos pontos fora indicando que não se adéquam para inferência. Nestes termos optou-se por não avaliar o efeito conjunto das variáveis. Os detalhes podem ser vistos na Figura 3.1.

Para modelagem preditiva[3] considerou-se dentro da perspectiva estatística a Regressão Logística e, dentro do campo de aprendizagem de máquinas, *Random Forest*. Tanto a Regressão Logística quanto *Random Forest* foram treinados com o uso da base de treino usando validação cruzada e ambas validadas na base de teste. Após seleção dos melhores modelos seguindo o critério da maximização da $AUC$, construiu-se uma nova a base de treinamento, conforme é recomendada para Modelos Aditivos Generalizados usando as predições das duas técnicas em teste e o vetor de supervisão, *status*. Da mesma forma foi construída uma nova base de validação, conforme indicado nas Figuras 3.2 e 3.3.

[1] Em geral modelos de Poisson com variância robusta são utilizados para dados de contagem, todavia na literatura da área de saúde aplica-se para estimar a Razão de Prevalência e, com esta perspectiva, foi adotado neste trabalho.

[2] Detalhamento nos itens: 3.1.1 Prevalência e 3.1.2 Razão de Prevalência.

[3] Resultados detalhados em 3.2 Abordagem preditiva.

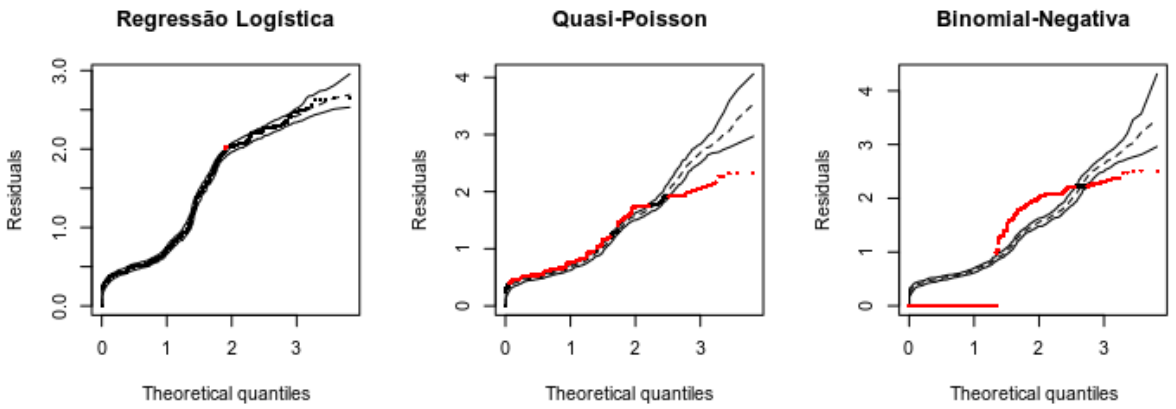


Figura 3.1: Envelopamento para regressão múltipla. Fonte: O autor.

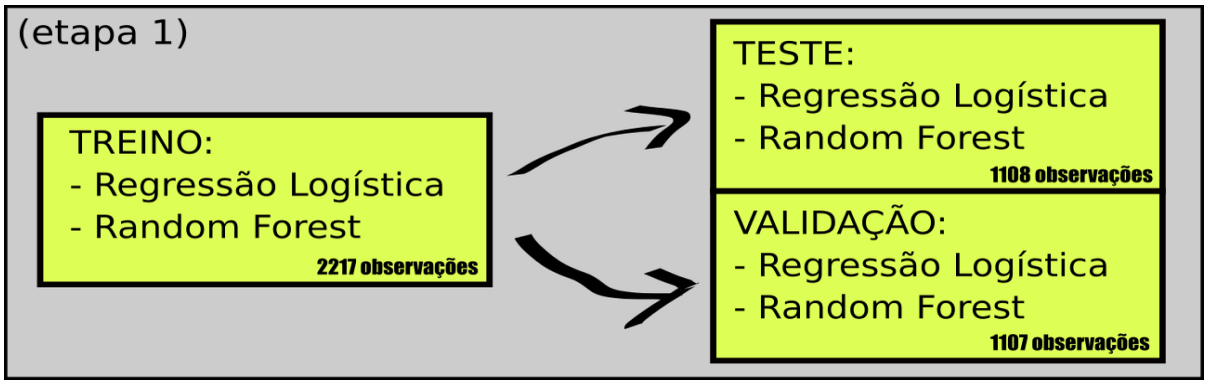


Figura 3.2: Construção da modelagem preditiva, primeira etapa. Fonte: O autor.

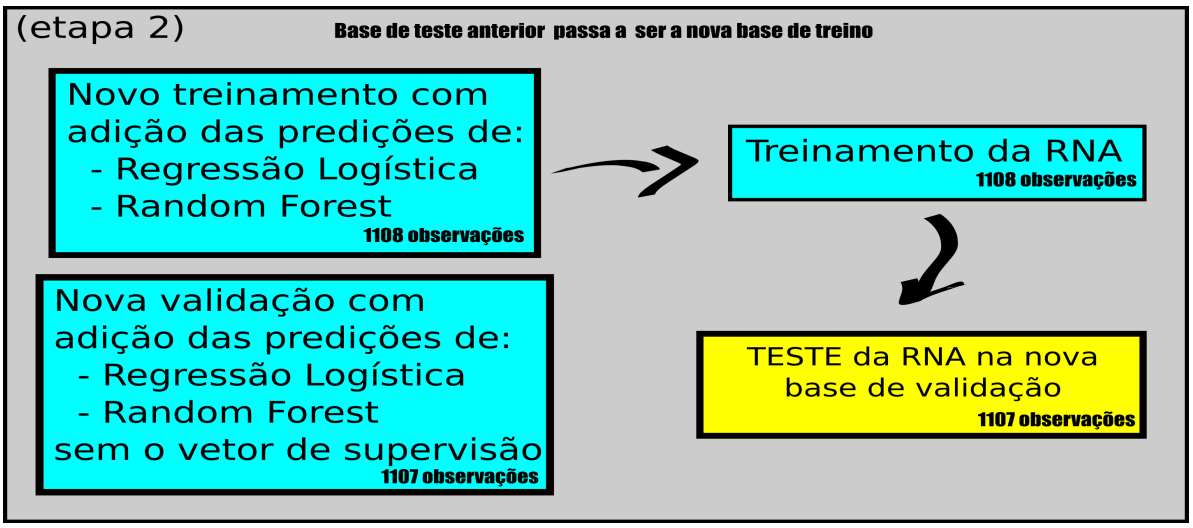


Figura 3.3: Construção da modelagem preditiva, segunda etapa. Fonte: O autor.

## 3.1 Abordagem explicativa

### 3.1.1 Prevalência

De acordo com a administração pública do IFB os fatores: sexo, grau de instrução, raça, estado civil, campus, forma de curso, modalidade de ensino, área do curso, origem escolar do aluno, necessidade específica, idade, motivo de escolha, quantidade de filhos e renda familiar são os mais importantes para avaliar a ausência de êxito discente.

Esta leitura segue avaliando as variáveis isoladamente. Nestes termos o sexo masculino tem uma prevalência maior (0.2018) do que o feminino (0.1652) indicando que mulheres estão mais propensas a se formar. Sobre ao grau de instrução, pessoas com superior completo tem alguma vantagem em relação aos outros níveis de instrução. Raça não apresenta tanta diferença e isso pode ser reflexo do público atendido. Estado civil sugere que pessoas separadas ou divorciadas têm prevalência menor do que as de união estável ou casadas.

Quanto ao *campus*, em específico o de Ceilândia (eixo de Ciências da Saúde) tem a menor prevalência de risco escolar (0.0632) enquanto o de Samambaia (eixo de Ciência Exatas e da Terra ou Engenharias) alarma com 0.4607. Note que isso pode ser motivado pela vocação do campus o que nos leva ao entendimento da área de ciências da saúde vantajosa se comparada as outras na percepção do cursista. Por outro lado os eixos tecnológicos do IFB foram constituídos mediante audiência pública com empresários que representam o arranjo produtivo local, ou seja, a população de cursistas não se mostra resiliente na área de exatas para atender a demanda da oferta de emprego. Cursos modulares sugerem que o formato de componentes, em que o aluno monta sua grade, traz potencial prejuízo nos indicadores da instituição, uma vez que a prevalência do grupo de risco é maior (0.209 contra 0.178).

A modalidade (Modal) semi-presencial (0.0757) se mostra mais eficiente do que a presencial (0.1531) ou à distância (0.208). Isso um ponto a ser considerado pelo Núcleo Docente Estruturante na adequação periódica dos planos de curso. Quando a origem escolar não importa se o estudante veio da escola pública ou particular e da mesma forma interpretamos o efeito da idade. Pessoas com necessidades específicas (0.2778) precisam de melhor acompanhamento para obterem êxito independente de se comparar aos que não têm necessidades específicas (0.180).

Quanto ao motivo de escolha, estudantes em que o curso não tem relação com a atuação profissional aparentam melhores condições de êxito (0.1475) se comparadas com estudantes que estão no curso para complementação (0.1899) ou já atuam na área (0.1853). Avaliando a quantidade de filhos, quem tem dois, está melhor ajustado ao êxito (0.1845) se comparado aos outros. Por fim, quem tem a renda familiar de até um salário mínimo

mostra uma barreira para se formar (0.2684) e a medida que a renda familiar aumenta, o êxito melhora. Para detalhamento veja o apêndice A: Tabelas - Prevalência partes I e II.

### 3.1.2 Razão de Prevalência

Após leitura das variáveis individualmente, o próximo passo natural é avaliar qual razão de prevalência é significativa dado o nível de referência. Por exemplo, cursos com poucos alunos, ou de informática, infraestrutura escolar, móveis, química e vestuário têm prevalência significativa tomando como base o de administração. Assim percebemos que o curso de móveis chega a ter 6.93 vezes a prevalência do curso de administração. Como segundo caso mais grave temos o curso de química com o multiplicador de 4.21.

Quanto ao estado civil a diferença entre solteiros e as outras classes não é significativa. O grau de instrução, superior completo, ainda que inconclusivo na inferência, mostra uma grande diferença qualitativa em relação ao estudante de nível médio (0,32). Para quem tem curso superior incompleto a razão de prevalência é praticamente a mesma do médio (0,96). Raça não mostra diferença significativa. Sexo feminino (0.82) indica que homens chegam a 22% a mais do que as mulheres. Para fins de inferência considerou-se o pvalor ajustado[4]. O detalhamento pode ser encontrado no apêndice A: Tabelas - Razão de Prevalência.

Por fim, na perspectiva da administração, a leitura de Prevalência e da Razão de Prevalência satisfaz com as informações extraídas o que uma coordenação de curso precisa. Portanto encerramos a leitura explicativa e partimos para a modelagem preditiva com o intuito de elaboração de ações preventivas para amparar o estudante quando der início ao seu curso.

## 3.2 Abordagem preditiva

Nesta parte detalharemos os resultados da modelagem preditiva por Regressão Logística, Random Forest e Redes Neurais Artificiais, conforme indicado nas Figuras 3.1 e 3.2 vistos anteriormente. A base de dados foi dividida nos subsets: treino (50%), teste (25%) e valid (25%), o que representa 2217, 1108 e 1107 observações respectivamente.

A construção dos modelos seguiram as etapas:

1. Treinamento usando Regressão Logística e Random Forest, com o subset treino.
    (a) Considerou-se a maximização do escore F1[5].

[4]O pvalor ajustado pode ser interpretado como o menor nível de significância em que a comparação específica será significativa se avaliada como teste de comparação múltipla.

[5]Escore F1 é um balanceamento descrito pela média harmônica entre Sensibilidade e Especificidade. 1 significa Sensibilidade e Especificidade perfeitos enquanto 0 é o pior resultado.

(b) *Early Stopping*[6].

2. Construção da base de teste2 e valid2 usando o rótulo de aprendizagem *status* e as predições dos modelos criados[7].

3. Treinamento da Rede Neural com teste2 e respectiva validação em valid2.

   (c) Considerou-se a maximização da AUC.

   (d) *Early Stopping*.

4. Foram utilizadas as bibliotecas *Caret*[22] e *H2O*[23] do R[24].

### 3.2.1 Regressão Logística e Random Forest

Para avaliar se os modelos treinados estão em concordância para as predições da base de teste e validação, optou-se pela realização das predições em ambas as bases e, para as probabilidades, temos uma suavização por curva de núcleo[8] e como resultado para Regressão Logística, temos um grau de explicação de 0.99, ou seja, para amostras independentes a predição é concordante, levando a validação do modelo. Da mesma forma a validação do Random Forest foi avaliada, chegando no $R^2$ de 0.78. A concordância do RF é relevante, porém menor do que RL. Esses resultados indicam que as predições em dados externos são mais semelhantes com a RL do que com RF.

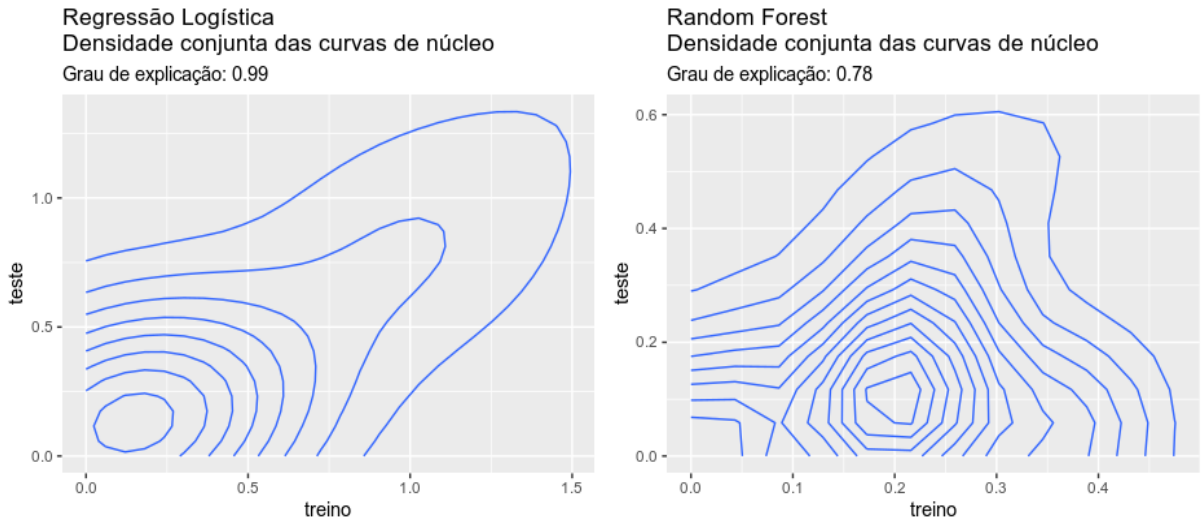


Figura 3.4: Validação dos modelos. Fonte: O autor.

[6] *Early Stopping* é um critério de parada preventiva de superajuste do modelo.

[7] Será tratado no tópico 3.2.2 Modelos Aditivos Generalizados.

[8] Como estamos interessados em comparar as distribuições preditas, a suavização por curva de núcleo serve como uma régua em que as densidades ficam pareadas.

Considerando a base de treino, RF se destaca com acurácia de 94% enquanto RL chega em 81%, da mesma forma temos o escore F1 com 0.87 e 0.46 respectivamente. Como estamos interessados em classificar a categoria 'sem_exito' então a Sensibilidade passa a ser um ponto importante na análise, levando o RF novamente em destaque com 0.77 contra 0.48 da RL. Avaliando a curva ROC e sua área, RF detém $AUC = 0.97$ e RL $AUC = 0.67$. Então a conclusão parcial é a que RF aprende um pouco melhor sobre as condições da classe de interesse, mas pode ser que esteja apenas memorizando as respostas do livro. Mais informações na Tabela 3.1 e na Figura 3.4.

| termo | est.RL | CI.low.RL | CI.high.RL | est.RF | CI.low.RF | CI.high.RF |
|---|---|---|---|---|---|---|
| accuracy | 0,81 | 0,79 | 0,82 | 0,94 | 0,93 | 0,95 |
| sensitivity | 0,48 | - | - | 0,77 | - | - |
| specificity | 0,88 | - | - | 1,00 | - | - |
| pos_pred_value | 0,45 | - | - | 1,00 | - | - |
| neg_pred_value | 0,89 | - | - | 0,93 | - | - |
| precision | 0,45 | - | - | 1,00 | - | - |
| recall | 0,48 | - | - | 0,77 | - | - |
| f1 | 0,46 | - | - | 0,87 | - | - |

Tabela 3.1: Escores dos modelos em treino, com o ponto de corte $c = 0.5$.

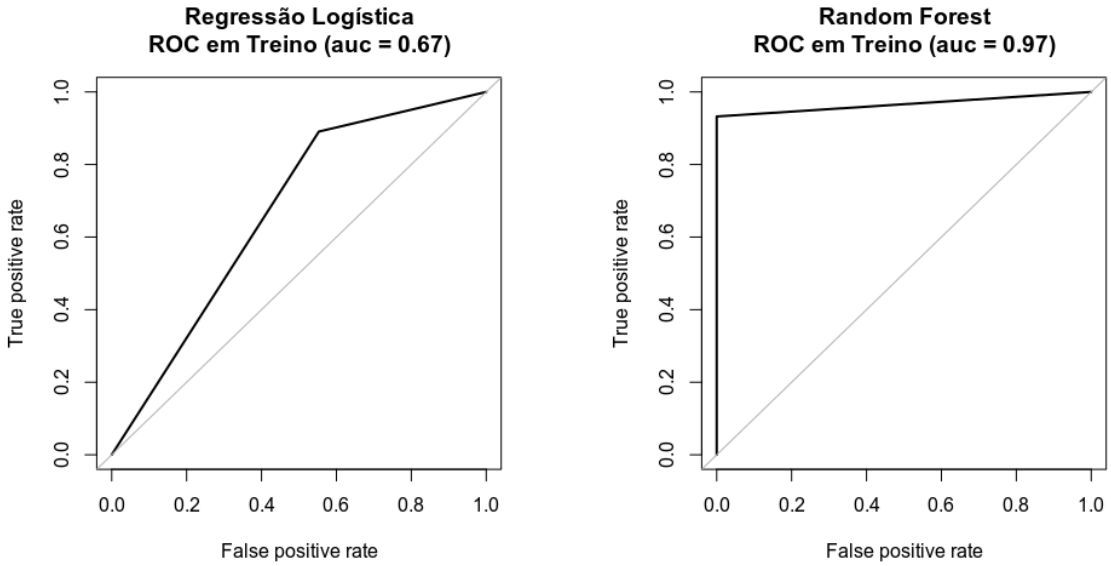


Figura 3.5: Curva ROC e AUC em treino. Fonte: O autor.

Considerando as bases de teste e validação a RL tem acurácia de 80%, sensitivity de 0.47, F1 de 0.44 e $AUC = 0.65$, se mostrando em geral melhor do que RF, que apresenta 71% de acurácia, sensitivity de 0.32, F1 de 0.40 e $AUC = 0.63$ em teste. A mesma leitura

segue para a base valid, com: 80%, 0.46, 0.45 e AUC de 0.66 na RL contra 73%, 0.36, 0.45 e 0.68 com RF, respectivamente.

| termo | est.RL | CI.low.RL | CI.high.RL | est.RF | CI.low.RF | CI.high.RF |
|---|---|---|---|---|---|---|
| accuracy | 0,80 | 0,78 | 0,83 | 0,71 | 0,68 | 0,73 |
| sensitivity | 0,47 | - | - | 0,32 | - | - |
| specificity | 0,87 | - | - | 0,87 | - | - |
| pos_pred_value | 0,41 | - | - | 0,52 | - | - |
| neg_pred_value | 0,90 | - | - | 0,75 | - | - |
| precision | 0,41 | - | - | 0,52 | - | - |
| recall | 0,47 | - | - | 0,32 | - | - |
| f1 | 0,44 | - | - | 0,40 | - | - |

Tabela 3.2: Escores dos modelos em teste.

| termo | est.RL | CI.low.RL | CI.high.RL | est.RF | CI.low.RF | CI.high.RF |
|---|---|---|---|---|---|---|
| accuracy | 0,80 | 0,77 | 0,82 | 0,73 | 0,7 | 0,76 |
| sensitivity | 0,46 | - | - | 0,36 | - | - |
| specificity | 0,87 | - | - | 0,89 | - | - |
| pos_pred_value | 0,44 | - | - | 0,60 | - | - |
| neg_pred_value | 0,88 | - | - | 0,76 | - | - |
| precision | 0,44 | - | - | 0,60 | - | - |
| recall | 0,46 | - | - | 0,36 | - | - |
| f1 | 0,45 | - | - | 0,45 | - | - |

Tabela 3.3: Escores dos modelos em valid.

Para avaliar se o resultado é coerente, realisou-se uma estimativa da distribuição da média da AUC por *bootstrap* dos modelos nas bases teste e valid, com 100 replicações. Como resultados temos valores semelhantes para treino, sendo $CI_{RL}(95\%) = [0.61, 0.68]$ e $CI_{RF}(95\%) = [0.61, 0.67]$ e para teste, $CI_{RL}(95\%) = [0.63, 0.69]$ e $CI_{RF}(95\%) = [0.66, 0.72]$.

Note que no *bootstrap* o processo é reamostrado, portanto existe um pequeno vício e os resultados não são independentes, deste modo uma análise de variância não é cabível para concluir se existe diferença entre as distribuições. Portanto optou-se construir a distribuição da diferença das AUCs dos modelos para teste e valid via *bootstrap*.

O resultado é um $CI_{RL-RF}(95\%) = [-0.0241, 0.0210]$ simétrico e, como o valor zero pertence ao intervalo, então não existe diferença entre os modelos quanto a AUC. Isso indica que ambos os modelos são coerentes para dados externos, uma vez que AUC não difere entre eles. Figuras 3.6 e 3.7.

Isso mostra claramente que a informação acurácia não deve ser o único fator para desenvolvimento, pois neste problema as classes seguem proporções desbalanceadas[9] e

[9]sem_exito (0.19), exito (0.81).

pode ocorrer do modelo oferecer uma acurácia alta com Sensibilidade pequena, gerando um falso negativo para pessoas em risco escolar, o que é prejudicial em ações preventivas.

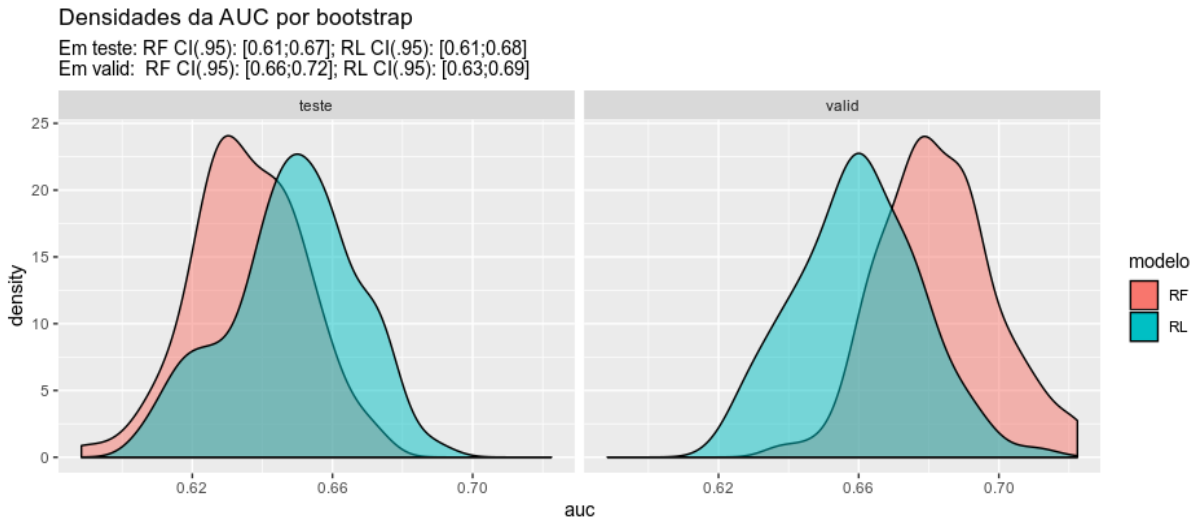


Figura 3.6: AUC bootstrap completo. Fonte: O autor.

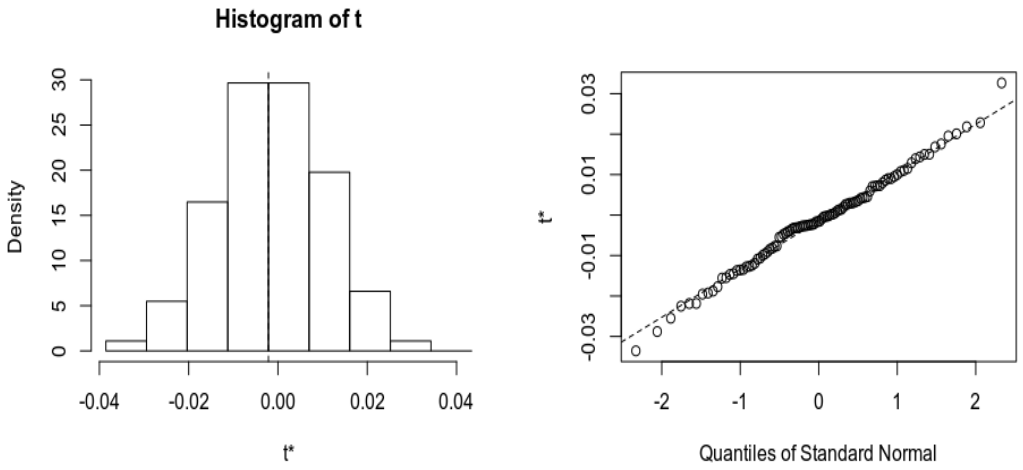


Figura 3.7: AUC bootstrap: Diferença entre modelos. Fonte: O autor.

Todavia AUC oferece o melhor caminho para avaliação do modelo, mas repare que ainda temos uma AUC pouco informativa tando para RL quanto para RF, sendo o ideal algo superior a 0.7[10] para se ter algum poder de discriminação. Mais informações estão nas tabelas: Escores dos modelos em teste, Escores dos modelos em valid, e, na figura: Curva ROC e AUC em teste e valid.

[10]Valor recomendado para estudos na área de humanas: [0.90,1.00] excelente (A); [0.80,0.90) bom (B); [0.70,0.80) justo (C); [0.60,0.70) pobre (D); [0.50,0.60) falho (F). Fonte: http://gim.unmc.edu/dxtests/roc3.htm.

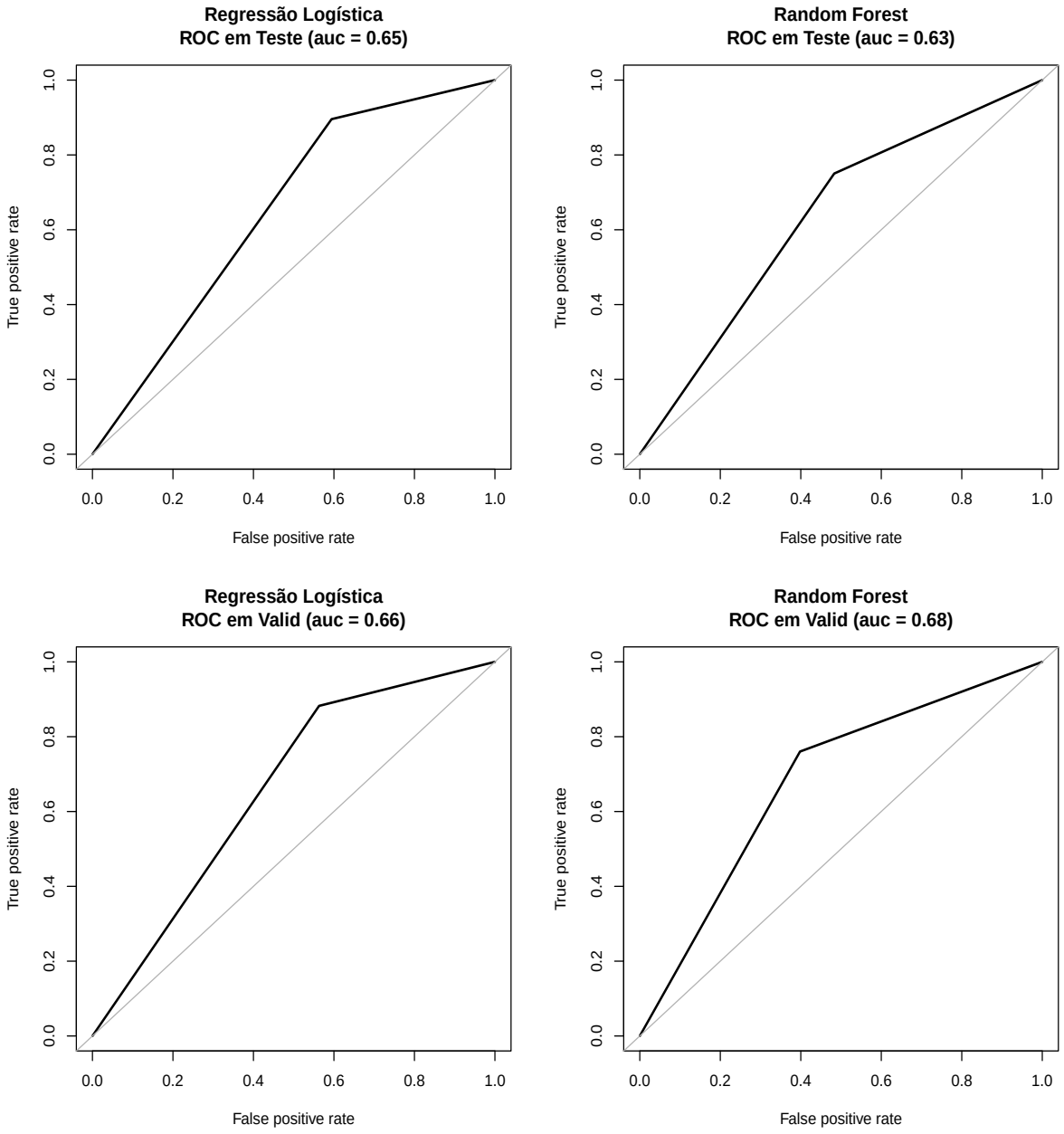


Figura 3.8: Curva ROC e AUC em teste e valid. Fonte: O autor.

### 3.2.2 Redes Neurais Artificiais

Algo que pode ser praticado com a finalidade de melhoria na AUC é o uso de Redes Neurais Artificiais que podem ser treinados com uma nova base, teste2, contendo apenas as predições dos modelos e o vetor *status*[11]. Considerando as predições dos modelos, tanto na base de teste quanto em valid, o grau de associação linear entre os resultados é significativo e moderado ($R^2 = 0.47$). Por outro lado RNAs não estão limitadas em relação ao efeito de multicolinearidade[25].

Figura 3.9: Comparativo dos modelos. Fonte: O autor.

Nestes termos foi construída uma RNA com 4 camadas ocultas de 39 neurônios[26] ativados por *Rectifier*[12], com a métrica de parada AUC máxima, seguindo a função de

[11] Considera-se essa construção como modelo combinado

[12] $f(x) = max\{0, x\}$, $x$ entrada do neurônio.

perda *Cross-Entropy*[13]. Foram usadas apenas 100 épocas e *Early Stopping*.

Retomando a perspectiva da administração dos *campi* a RNA traz uma melhora significativa ao avaliar a classificação correta do grupo de interesse, sendo avaliada pelas predições dentro da linha sem_exito, da matriz de confusão, encontramos a RNA com 0.6748, bem superior a RL (0.4369) e RF (0.6019). A AUC também mostra uma avanço, chegando a 0.73 contra 0.66 da RL e 0.68 da RF, que passa a ser significativo na perspectiva de humanas aplicadas.

**Regressão Logística**

```
Confusion Matrix and Statistics

          Reference
Prediction  sem_exito exito
  sem_exito        90   116
  exito           106   795

               Accuracy : 0.7995
                 95% CI : (0.7746, 0.8227)
    No Information Rate : 0.8229
    P-Value [Acc > NIR] : 0.9803

                  Kappa : 0.3253

 Mcnemar's Test P-Value : 0.5458

            Sensitivity : 0.4592
            Specificity : 0.8727
         Pos Pred Value : 0.4369
         Neg Pred Value : 0.8824
             Prevalence : 0.1771
         Detection Rate : 0.0813
   Detection Prevalence : 0.1861
      Balanced Accuracy : 0.6659

       'Positive' Class : sem_exito
```

**Random Forest**

```
Confusion Matrix and Statistics

          Reference
Prediction  sem_exito exito
  sem_exito       124    82
  exito           216   685

               Accuracy : 0.7308
                 95% CI : (0.7036, 0.7567)
    No Information Rate : 0.6929
    P-Value [Acc > NIR] : 0.003142

                  Kappa : 0.2896

 Mcnemar's Test P-Value : 1.314e-14

            Sensitivity : 0.3647
            Specificity : 0.8931
         Pos Pred Value : 0.6019
         Neg Pred Value : 0.7603
             Prevalence : 0.3071
         Detection Rate : 0.1120
   Detection Prevalence : 0.1861
      Balanced Accuracy : 0.6289

       'Positive' Class : sem_exito
```

**Rede Neural Artificial**

```
Confusion Matrix and Statistics

          Reference
Prediction  sem_exito exito
  sem_exito       139    67
  exito           195   706

               Accuracy : 0.7633
                 95% CI : (0.7372, 0.7881)
    No Information Rate : 0.6983
    P-Value [Acc > NIR] : 8.509e-07

                  Kappa : 0.3697

 Mcnemar's Test P-Value : 4.292e-15

            Sensitivity : 0.4162
            Specificity : 0.9133
         Pos Pred Value : 0.6748
         Neg Pred Value : 0.7836
             Prevalence : 0.3017
         Detection Rate : 0.1256
   Detection Prevalence : 0.1861
      Balanced Accuracy : 0.6647

       'Positive' Class : sem_exito
```

Figura 3.10: Matriz de confusão: Comparativo da RNA com RL e RF em valid2. *Output* do comando *caret* :: $confusionMatrix()$. Fonte: O autor.

Uma pergunta no mínimo razoável para este fechamento do trabalho é se este esforço para criar a RNA valeu a pena. Para responder recorreremos novamente ao *bootstrap*, com 100 replicações, avaliando a distribuição da diferença entre AUC da RNA para a RL e RF. O resultado é animador contando com $CI_{RN-RL}(95\%) = [0.0455, 0.0993]$, $CI_{RN-RF}(95\%) = [0.0219, 0.0828]$ e $CI_{RN}(95\%) = [0.6919, 0.7625]$, mostrando-se superior em desempenho para AUC. Detalhamento na Figura 3.12.

[13]Calcula a entropia total entre distribuições. Tem origem na Teoria da Informação. $H(p, q) = H(p) + D_{KL}(p||q)$.

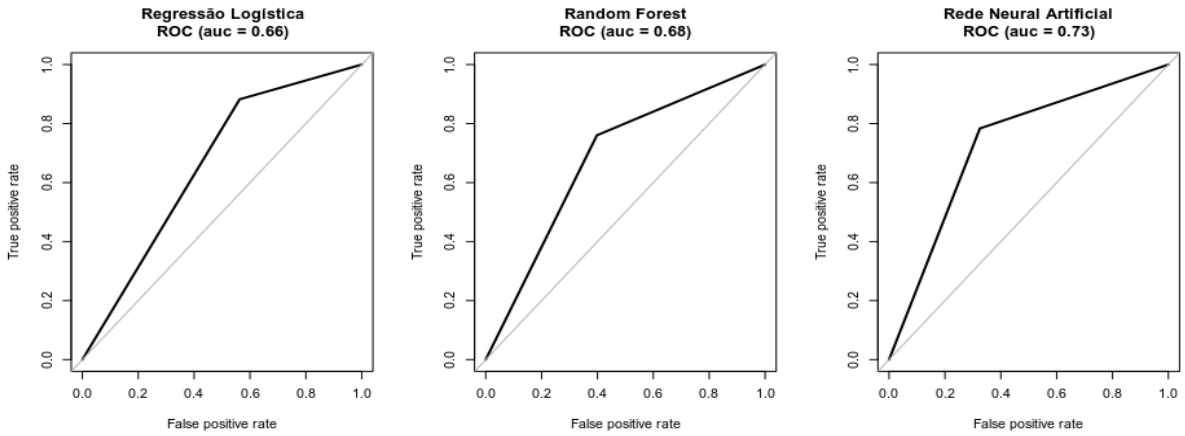


Figura 3.11: ROC e AUC: Comparativo da RNA com RL e RF em valid2. Fonte: O autor.

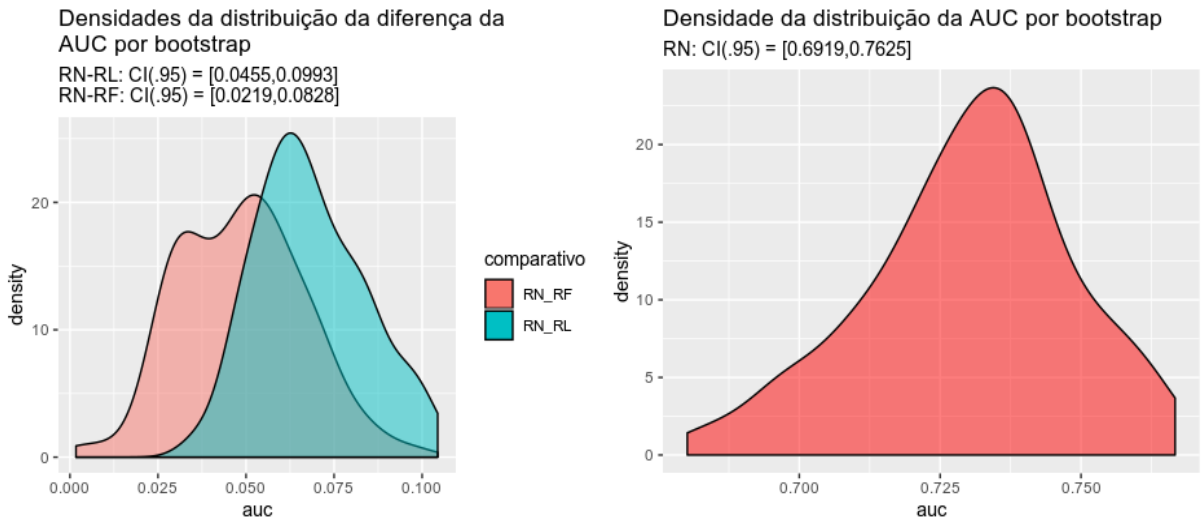


Figura 3.12: Desempenho da RNA: AUC em valid2, comparanto com RL e RF. Fonte: O autor.

# 4 Conclusão

Considerando a temática evasão escolar, este trabalho resume alguns aspectos da Economia Educacional, indicando dentro da ralidade do Instituto Federal de Brasília, que historicamente (2014 a 2017) 23% são considerados grupos de risco escolar, definido por cancelado, evadido, trancado ou integralizado. Focando apenas nos cursos técnicos essa proporção cai para 19%.

Os estudos qualitativos foram direcionados seguindo o interesse da administração dos *campi*. Para tanto foi conduzido por regressão de Poisson com variância robusta, a busca dos efeitos explicativos das variáveis disponíveis no questionário sócio-econômico em relação ao risco do estudante. Percebe-se que homens têm, em relação ao Risco Escolar, prevalência maior (20.18%) do que as mulheres (16.52%), além deste concluímos que cursos na área de saúde são benéficos para o indicador Relação Professor Aluno (RAP) que é a proporção de discentes por docente, uma vez que a prevalência é pequena (6.32%) ao ser comparada com ciências exatas e da terra ou engenharias (46.07%). A modalidade também é fator importante para a RAP mostrando que o semi-presencial (7.57%) é mais eficiente que outras modalidades (chegando até em 20.8%). Portadores de necessidades específicas estão em maior prevalência de Risco Escolar (27.78%) em relação a quem não tem necessidades específicas(18.99%). Quanto a Razão de Prevalência (RP) a área do curso, Técnico em Produção de Móveis e Técnico em Química), se destacam com RP de 6.93 e 4.21, respectivamente ao serem comparadas com Técnico em Administração.

Avaliando a modelagem preditiva, os dados foram separados em três partições, treino (50%), teste (25%), valid (25%), sendo concluída em duas etapas. Na primeira, em treino, usou-se Regressão Logística (RL) e Random Forest (RF) maximizando F1 considerando o critério de *Early Stopping* para evitar *overfit*. Os modelos chegaram em 80% e 71% de acurácia e F1 de 0.44 e 0.4, para RL e RF respectivamente. AUC foi de 0.65 (RL) e 0.63 (RF), que é baixo para fins de predição em humanas aplicadas. Avaliando se os modelos têm desempenho da AUC semelhante por *bootstrap*, concluímos que não é significativa uma vez que o intervalo de confiança percentílico da diferença de AUC a 95% entre os modelos é $[-0.0241, 0.0210]$.

Para melhorar o resultado da AUC usou-se as predições dos modelos em treino, construindo uma nova base contendo as predições e o vetor de resposta *status*, teste2. O mesmo foi realizado em valid criando valid2. Assim aplicou-se a Rede Neural Artificial, com quatro camadas ocultas contendo 39 neurônios cada, com 100 épocas e *Early Stopping*, em teste2 para fins de treinamento e posterior validação em valid2.

A AUC da RNA foi de 0.73, agora aceitável e superior a da RL (0.66) e RF (0.68), desta forma ocorreu o benefício esperado que é principalmente evitar problemas de classificação dos sem êxito, ou seja, a taxa de verdadeiros positivos passa a ser 0.6748 contra 0.4369 (RL) e 0.6019 (RF). O modelo da RNA foi comparado por *bootstrap* ao RL e RF gerando ICs percentílicos de $[0.0455, 0.0993]$ (RL) e $[0.0219, 0.0828]$ (RF) a 95% de cobertura. O desempenho da AUC, simulado pela mesma técnica, da RNA é de $[0.6919, 0.7625]$.

# Referências


[1] MARIANO, A. M., ROCHA M. S.: *Revisão da literatura: Apresentação de uma abordagem integradora.* AEDEM International Conference, 2017. 1

[2] BORDIEU, Pierre.: *The forms of capital.* Handbook of Theory of Research for the Sociology of Education. pp. 46-86, 1986. 2

[3] BASTRO, Cláudio de Moura.: *Investimento em educação no brasil: Comparação de três estudos.* Pesquisa e Planejamento Econômico (PPE). http://ppe.ipea.gov.br, 1971. 3

[4] PEREIRA, Mirian Tomiato; LOPES, Janete Leige.: *A importância do capital humano para o crescimento econômico.* IX EPCT Encontro de Produção Científica e Tecnológica Campo Mourão, 27 a 31 de Outubro de 2014 ISSN 1981-64800., 2014. 4

[5] BONELLI, Regis; RAMOS, Lauro.: *Distribuição de renda no brasil: Avaliação das tendência de longo prazo e mudanças na desigualdade desde os meados dos anos 70.* Repositório do IPEA. http://repositorio.ipea.gov.br/bitstream/11058/2493/1/td_0288.pdf, 1993. 4

[6] IBGE: *Pesquisa nacional por amostra de domicílios 1992/2007.* Repositório de Séries Estatísticas. https://seriesestatisticas.ibge.gov.br, 2019. 4

[7] IFB, Instituto Federal de Brasília.: *Missão, visão e valores.* Site institucional. https://www.ifb.edu.br/institucional/missao, 2019. 5

[8] SORENSEN, L. C.: *Big data in educational administration: An application for predicting school dropout risk.* Educational Administration Quarterly., 55, 2019. 6

[9] Erikson, G. e L. Walkthrough. Franks: *Develop a predictive analytics solution for credit risk assessment in azure machine learning.* Site: https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-walkthrough-develop-predictive-solution, 2017. 10

[10] AX., Microsoft Dynamics: *Use machine learning - based microsoft cognitive services for product recommendations.* Site: https://mbs.microsoft.com/Files/public/CS/AX2012R3/WhitepaperCortanaintelligenceservicesmanagementinAX2012R3.pdf, 2012. 10

[11] Nimeroff, J.: *How machine learning will be used for marketing in 2017.* Site: https://www.forbes.com/sites/forbestechcouncil/2017/03/10/how-machine-learning-will-be-used-for-marketing-in-2017/5ad950ea6d3d, 2017. 10

[12] Hill, K.: *How target figured out a teen girl was pregnant before her father did.* Site: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-fa ther-did/6b6f03b26668, 2012. 10

[13] Lehman, Li-wei, Roger Mark e Shamim Nemati.: *A model-based machine learning approach to probing autonomic regulation from nonstationary vital-signs time series.* IEEE journal of biomedical and health informatics, 2017. 10

[14] Alexander, James, et al.: *A novel methodology for in vivo endoscopic phenotyping of colorectal cancer based on real-time analysis of the mucosal lipidome: a prospective observational study of the knife.* Surgical endoscopy. pp 1361-1370, 31(3), 2017. 10

[15] JOHNSON, Richard A. e Dean W. WICHERN: *Applied multivariate statistical analysis.* 6nd Edition. Pearson Prentice Hall., 2007. 10

[16] Le, Laetitia, et al.: *Machine learning for classification and quantification of monoclonal antibody preparations for cancer therapy.* arXiv preprint arXiv:1705.07099., 2017. 10

[17] AGRESTI, Alan.: *An introduction to categorical data analysis.*, volume 55. 2019. 12

[18] ZOU, Guangyong.: *A modified poisson regression approach to prospective studies with binary data.* American Journal of Epidemiology. pp 404–446, 159(7), 2019. 14

[19] Yu-Wei, Chiu.: *Machine learning with r cookbook.* Packt Publishing., 2015. 14

[20] Breiman, Leo.: *Random forests.* Kluwer Academic Publishers., 2001. 15

[21] Hastie, Trevor, et al.: *The elements of statistical learning data mining, inference, and prediction.* Springer., 2001. 15

[22] Kuhn, Max: *caret: Classification and Regression Training*, 2020. `https://CRAN.R-project.org/package=caret`, R package version 6.0-86. 22

[23] LeDell, Erin, Navdeep Gill, Spencer Aiello, Anqi Fu, Arno Candel, Cliff Click, Tom Kraljevic, Tomas Nykodym, Patrick Aboyoun, Michal Kurka e Michal Malohlava: *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*, 2020. `https://CRAN.R-project.org/package=h2o`, R package version 3.30.0.1. 22

[24] R Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. `https://www.R-project.org/`. 22

[25] Rocha, Samille, et al.: *Uso de redes neurais para previsão de produção de viagens: Uma análise agregada.* XXIX ANPET., 2015. 27

[26] Masters, Timothy.: *Practical neural network recipes in c++.* Morgan Kaufmann., 1993. 27

# A : Tabelas

| Variável | Grupo | Frequência | Total | Proporção | IC(95%) | |
|---|---|---|---|---|---|---|
| | | | | | Inf. | Sup. |
| Sexo | Masculino | 404 | 2002 | 0,2018 | 0,1842 | 0,2194 |
| | Feminino | 430 | 2603 | 0,1652 | 0,1509 | 0,1795 |
| Grau de instrução | Ensino Médio Completo | 789 | 4269 | 0,1848 | 0,1732 | 0,1965 |
| | Superior Incompleto | 31 | 175 | 0,1771 | 0,1206 | 0,2337 |
| | Pos+ | 8 | 61 | 0,1311 | 0,0464 | 0,2159 |
| | Superior Completo | 6 | 100 | 0,06 | 0,0135 | 0,1065 |
| Raça | Branca | 217 | 1088 | 0,1994 | 0,1757 | 0,2232 |
| | Parda | 429 | 2382 | 0,1801 | 0,1647 | 0,1955 |
| | Preta | 131 | 765 | 0,1712 | 0,1445 | 0,1979 |
| | Indígena ou Amarela | 24 | 165 | 0,1455 | 0,0917 | 0,1992 |
| Estado civil | Viúvo(a) | 5 | 22 | 0,2273 | 0,0522 | 0,4024 |
| | UE ou Casado(a) | 280 | 1395 | 0,2007 | 0,1797 | 0,2217 |
| | Solteiro(a) | 494 | 2845 | 0,1736 | 0,1597 | 0,1876 |
| | Separado(a) ou Divorciado(a) | 30 | 195 | 0,1538 | 0,1032 | 0,2045 |
| Campus | CSAM | 170 | 369 | 0,4607 | 0,4098 | 0,5116 |
| | CTAG | 190 | 690 | 0,2754 | 0,242 | 0,3087 |
| | CTGC | 34 | 127 | 0,2677 | 0,1907 | 0,3447 |
| | CRFI | 84 | 461 | 0,1822 | 0,147 | 0,2175 |
| | CGAM | 114 | 709 | 0,1608 | 0,1338 | 0,1878 |
| | CPLA | 138 | 1023 | 0,1349 | 0,114 | 0,1558 |
| | CSSE | 39 | 419 | 0,0931 | 0,0653 | 0,1209 |
| | CBRA | 48 | 519 | 0,0925 | 0,0676 | 0,1174 |
| | CCEI | 17 | 269 | 0,0632 | 0,0341 | 0,0923 |
| Forma de curso | Componente ou Outros | 90 | 430 | 0,2093 | 0,1709 | 0,2478 |
| | Modular | 744 | 4175 | 0,1782 | 0,1666 | 0,1898 |
| Modal | Presencial | 562 | 2702 | 0,208 | 0,1927 | 0,2233 |
| | À distância | 253 | 1652 | 0,1531 | 0,1358 | 0,1705 |
| | Semi-Presencial | 19 | 251 | 0,0757 | 0,043 | 0,1084 |

Tabela A.1: Prevalência - parte I.

| Variável | Grupo | Frequência | Total | Proporção | IC(95%) | |
|---|---|---|---|---|---|---|
| | | | | | Inf. | Sup. |
| Área | Outros | 412 | 1618 | 0,2546 | 0,2334 | 0,2759 |
| | Ciência Exatas e da Terra ou Engenharias | 113 | 707 | 0,1598 | 0,1328 | 0,1868 |
| | Ciências Agrárias | 55 | 346 | 0,159 | 0,1204 | 0,1975 |
| | Ciências Sociais Aplicadas ou Ciência Humanas | 242 | 1795 | 0,1348 | 0,119 | 0,1506 |
| | Ciências da Saúde | 12 | 139 | 0,0863 | 0,0396 | 0,133 |
| Origem escolar | Pública | 561 | 3211 | 0,1747 | 0,1616 | 0,1878 |
| | Particular | 127 | 745 | 0,1705 | 0,1435 | 0,1975 |
| Necessidade específica | Alguma necessidade | 5 | 18 | 0,2778 | 0,0709 | 0,4847 |
| | Nenhuma necessidade | 829 | 4587 | 0,1807 | 0,1696 | 0,1919 |
| Idade | até 31 | 421 | 2308 | 0,1824 | 0,1667 | 0,1982 |
| | 32 ou mais | 413 | 2297 | 0,1798 | 0,1641 | 0,1955 |
| Motivo de escolha | Relação com Trabalho ou Complementação | 79 | 416 | 0,1899 | 0,1522 | 0,2276 |
| | Realização profissional ou Profissionalização | 189 | 1020 | 0,1853 | 0,1615 | 0,2091 |
| | Sem relação Profissional | 32 | 217 | 0,1475 | 0,1003 | 0,1946 |
| Quantidade de filhos | 0 | 201 | 903 | 0,2226 | 0,1955 | 0,2497 |
| | 1 | 59 | 249 | 0,2369 | 0,1841 | 0,2898 |
| | 2 | 43 | 233 | 0,1845 | 0,1347 | 0,2344 |
| | 3+ | 39 | 154 | 0,2532 | 0,1846 | 0,3219 |
| Renda familiar (nº salário mínimo) | Até 1 | 62 | 231 | 0,2684 | 0,2113 | 0,3255 |
| | Mais de 1 e menos que 2 | 145 | 685 | 0,2117 | 0,1811 | 0,2423 |
| | Mais que 2 e menos do que 5 | 105 | 461 | 0,2278 | 0,1895 | 0,266 |
| | 5 ou + | 30 | 162 | 0,1852 | 0,1254 | 0,245 |

Tabela A.2: Prevalência - parte II.

| Variável | Categoria | RP | 2.5 % | 97.5 % | pvalor | padj | Significativo |
|---|---|---|---|---|---|---|---|
| Curso | Cursos com poucos alunos | 2,67 | 1,63 | 4,26 | 0,0001 | 0,0013 | Sim |
| | Informatica | 3,54 | 2,49 | 5,08 | 0 | 0 | Sim |
| | Infraestrutura Escolar | 2,23 | 1,57 | 3,2 | 0 | 0,0002 | Sim |
| | Móveis | 6,93 | 5,1 | 9,59 | 0 | 0 | Sim |
| | Química | 4,21 | 2,73 | 6,43 | 0 | 0 | Sim |
| | Vestuário | 2,97 | 2,13 | 4,21 | 0 | 0 | Sim |
| | Seguranca | 0,53 | 0,3 | 0,9 | 0,0236 | 0,4729 | Não |
| | Servicos | 1,43 | 1,03 | 2 | 0,0351 | 0,6319 | Não |
| | Agro+ | 1,44 | 1 | 2,07 | 0,0502 | 0,8534 | Não |
| | Controle Ambiental | 1,14 | 0,6 | 2,02 | 0,6699 | 1 | Não |
| | Cozinha | 1,39 | 0,95 | 2,02 | 0,0874 | 1 | Não |
| | Equipamentos Biomédicos | 0,82 | 0,44 | 1,44 | 0,5135 | 1 | Não |
| | Logistica | 0,82 | 0,56 | 1,21 | 0,3175 | 1 | Não |
| | Meio Ambiente | 1,33 | 0,9 | 1,97 | 0,1513 | 1 | Não |
| | Multimeios Didaticos | 1,27 | 0,85 | 1,88 | 0,2463 | 1 | Não |
| | Secretariado | 1,15 | 0,7 | 1,84 | 0,5744 | 1 | Não |
| Estado Civil | Separado(a) ou Divorciado(a) | 0,89 | 0,62 | 1,22 | 0,4772 | 1 | Não |
| | UE ou Casado(a) | 1,16 | 1,01 | 1,32 | 0,0324 | 0,6162 | Não |
| | Viúvo(a) | 1,31 | 0,52 | 2,65 | 0,5084 | 1 | Não |
| Grau de instrução | Pos+ | 0,71 | 0,35 | 1,26 | 0,2863 | 1 | Não |
| | Superior Completo | 0,32 | 0,14 | 0,62 | 0,0024 | 0,0512 | Não |
| | Superior Incompleto | 0,96 | 0,68 | 1,31 | 0,798 | 1 | Não |
| Raça | Indígena ou Amarela | 0,73 | 0,49 | 1,05 | 0,1051 | 1 | Não |
| | Parda | 0,9 | 0,78 | 1,05 | 0,1762 | 1 | Não |
| | Preta | 0,86 | 0,7 | 1,04 | 0,1281 | 1 | Não |
| Sexo | Feminino | 0,82 | 0,72 | 0,93 | 0,0014 | 0,0314 | Sim |

Tabela A.3: Razão de Prevalência.