UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

EDUARDO GABRIEL PINHEIRO

COMBINAÇÃO LINEAR DE CLASSIFICADORES PARA ANÁLISE DE RISCO NA EVASÃO DA UFC

FORTALEZA

2018

EDUARDO GABRIEL PINHEIRO

COMBINAÇÃO LINEAR DE CLASSIFICADORES PARA ANÁLISE DE RISCO NA EVASÃO DA UFC

Monografia submetida à Coordenação do Curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. André Jalles Monteiro

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

P718c Pinheiro, Eduardo Gabriel.
Combinação Linear de Classificadores para Análise de Risco na Evasão da UFC / Eduardo Gabriel Pinheiro. – 2018.
63 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Estatística, Fortaleza, 2018.
Orientação: Prof. Dr. André Jalles Monteiro.

1. Classificadores. 2. Curva ROC. 3. AUC. 4. sensibilidade. 5. especificidade. I. Título.

CDD 519.5

EDUARDO GABRIEL PINHEIRO

COMBINAÇÃO LINEAR DE CLASSIFICADORES PARA ANÁLISE DE RISCO NA EVASÃO DA UFC

Monografia submetida à Coordenação do Curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Bacharel em Estatística.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. André Jalles Monteiro (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Rafael Bráz Azevedo Farias
Universidade Federal do Ceará (UFC)

Prof. Dr. Carlos Almir Monteiro de Holanda
Universidade Federal do Ceará (UFC)

À minha Família

AGRADECIMENTOS

Agradeço primeiramente a Deus pela oportunidade, força, saúde e coragem para enfrentar e superar meus desafios.

Aos meus pais Francisco de Castro Pinheiro e Jonailda Gabriel Pinheiro que sempre tiveram como prioridade o bem dos filhos, estando presentes para compartilhar dos sucessos e também consolar nos fracassos.

Ao meu irmão Felipe Gabriel Pinheiro que sempre me ajudou em todas as minhas dificuldades, o primeiro e melhor amigo da minha vida.

Ao Professor Dr. André Jalles Monteiro pelos conselhos profissionais e pessoais e orientação para a elaboração deste trabalho.

Ao CNPq por me auxiliar com a bolsa de Iniciação Científica.

Ao Professor Dr. João Maurício Araújo Mota, pelos ensinamentos sem preço, e imensa dedicação e preocupação com o aprendizado dos seus alunos.

Ao PET-estatística UFC e ao Professor Dr. Júlio Francisco Barros Neto pelos conselhos, apoio e por proporcionar oportunidades para que pudesse aprender e amadurecer profissionalmente e pessoalmente.

Ao Professor Dr. Ronald Targino Nojosa pelo apoio e contribuição no meu processo de aprendizagem.

Ao Professor Dr. Pushpa Narayan Rathie pelos ensinamentos e por partilhar um pouco de sua experiência.

Aos meus amigos Victor Máximo, Jonson, Luan, Ramon, Alisson, Liderson, Áurea, João Victor, Ramon Viana, Danrley, Allyson, Victor, Lucas (conhecido como Maraca), Diego, John e Roberto, pelos momentos de alegria e de convivência, estando presentes nos momentos de dificuldades e partilhando destes em alguns casos.

Ao Departamento de Estatística e Matemática Aplicada da UFC, por proporcionar oportunidades para a minha formação profissional.

Agradeço a todos que de algum modo me ajudaram a vencer os desafios que surgiram na vida.

"Compare yourself to who you were yesterday, not to who someone else is today."

(Jordan B. Peterson)

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

(Ronald Fisher)

RESUMO

Com a mudança do processo seletivo do Vestibular para o ENEM/SiSU na Universidade Federal do Ceará – UFC, o fenômeno da evasão discente se tornou mais precoce, acontecendo, em grande maioria, nos dois primeiros semestres do curso. Tal fenômeno acarreta prejuízos consideráveis para as instituições, assim como para a sociedade de modo geral. Desse modo, um estudo sobre essa problemática e a elaboração de mecanismos destinados a perceber o mais rápido possível a possibilidade da evasão se faz necessário. Uma das ações possíveis é a identificação dos discentes sujeitos à evasão com base em alguma métrica avaliativa, denotado aqui por classificador. Para a seleção de bons classificadores existem diferentes medidas destinadas a avaliar sua eficiência, como acurácia, sensibilidade, especificidade e a curva ROC (receiver operation characteristic). Esta que é bastante aplicada em diversas áreas do conhecimento. Dela obtêm-se diversas outras medidas, tais como: a AUC (área under curve) e pAUC(partial area under curve), tais medidas são fundamentadas por uma repertório teórico considerável. O objetivo do presente trabalho é abordar técnicas destinadas a combinação de diferentes classificadores a fim de formar uma melhor capacidade discriminativa. Aplicando tal ferramental em dados práticos relativos aos discentes ingressantes na UFC, nos cursos do Centro de Tecnologia, para uma verificação da capacidade desses métodos na percepção do risco à evasão destes discentes com a utilização das notas básicas do ENEM.

Palavras-chave: Classificadores, Curva ROC, AUC, sensibilidade, especificidade.

ABSTRACT

With the change from the selective process from the entrance exam to the ENEM / SiSU at the Federal University of Ceará - UFC, the phenomenon of student evasion became more precocious, occurring in the majority of the first two semesters of the course. This phenomenon causes considerable damage to the institutions as well as to society as a whole. In this way, a study on this problem and the elaboration of mechanisms to perceive the possibility of avoidance as quickly as possible is necessary. One of the possible actions is the identification of the students subject to evasion based on some evaluative metric, denoted here by classifier. For the selection of good classifiers, there are different measures to evaluate its efficiency, such as accuracy, sensitivity, specificity and the ROC (receiver operation characteristic) curve. This is quite applied in several areas of knowledge. From this, several other measures are obtained, such as the AUC (area under curve) and pAUC (partial area under curve), these measures are based on a considerable theoretical repertoire. The objective of the present work is to approach techniques aimed at combining different classifiers in order to form a better discriminative capacity. Applying this tool in practical data related to the students entering the UFC, in the courses of the Technology Center, to verify the ability of these methods in the perception of the evasion risk of these students with the use of basic grades of ENEM.

Keywords: Classifiers. ROC curve. AUC. sensitivity. specificity.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Métodos de classificação são utilizados em estudos práticos de diversas áreas com o objetivos de agrupar unidades observáveis com base em categorias de interesse. Na Biologia, por exemplo, o termo taxonomia refere-se à classificação de grupos biológicos com base em aspectos comuns; na Economia instituições financeiras utilizam modelos de classificação de risco em seus processos de concessão de crédito; na Medicina em estudos epidemiológicos, testes de diagnóstico são feitos com finalidade de identificar algumas doenças; etc.

Uma das formas possíveis de classificação pode ser viabilizada por intermédio de uma variável numérica, em que, neste caso, a classificação se dá por intermédio de intervalos, com a utilização de diversos pontos de corte, ou por intermédio de um ponto de corte único, em que até um específico valor de referência há um tipo de classificação e a partir deste valor uma outra classificação.

Atualmente no ensino superior existe uma problemática com relação à evasão discente que tem ocorrido principalmente no primeiro ano da graduação (LOBO, 2017), esse fenômeno provoca sérios prejuízos para as instituições, necessitando de atenção para a compreensão de suas causas. Uma alternativa para lidar com essa situação pode ser tratá-lo como um problema de classificação, com a adoção de uma métrica capaz de auxiliar na identificação de alunos propícios a desistência.

A presente monografia tem por objetivo apresentar propostas com base na curva ROC que serão utilizadas no estudo de percepção da evasão na UFC. As métricas utilizadas serão as notas obtidas pelos alunos no ENEM, em seus respectivos processos seletivos, no momento de ingresso na Universidade. Essa combinação linear assim como o processo de escolha de pontos de corte para a classificação podem ser utilizadas na escolha de pesos nas respectivas provas do ENEM, assim como na definição de notas mínimas para o ingresso como previsto na PORTARIA NORMATIVA Nº 21, DE 5 DE NOVEMBRO DE 2012, que dispõe sobre o Sistema de Seleção Unificada - SiSU.

Para uma melhor efetividade na percepção da evasão discente, além da verificada nos primeiros semestres, foram considerados os alunos ingressantes nos cursos do Centro de Tecnologia, por intermédio do processo ENEM/SiSU, nos anos de 2014 e 2015, observando seus respectivos status de matrícula no segundo semestre do ano de 2018.

Quanto à estrutura de elaboração desta monografia, adotou-se o seguinte critério: o primeiro capítulo, a introdução, apresenta a proposição do tema com seus objetivos, justificativa,

percurso metodológico e estrutura. No Capítulo 2 será apresentado a fundamentação teórica do trabalho, detalhando formalmente o processo de classificação, medidas de acurácia, bem como propriedades envolvendo curva ROC e medidas associadas à mesma. No Capítulo 3 são apresentados e discutidos diferentes métodos de combinação linear com base na curva ROC. No Capítulo 4 será realizada a aplicação referente ao estudo da evasão. No Capítulo 5, serão feitas as considerações finais do trabalho e propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Seja $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N\}$ elementos de uma população Ω e T uma variável dicotômica que caracteriza os elementos populacionais em duas categorias possíveis d_0 ou d_1; Ω_0 e Ω_1 são partições de Ω, tal que $\{T(\varepsilon) = d_0 \rightarrow \varepsilon \in \Omega_0\}$, $\{T(\varepsilon) = d_1 \rightarrow \varepsilon \in \Omega_1\}$. Existem situações em que não é possível obter $T(\varepsilon)$, impossibilitando identificar quais elementos pertecem a Ω_0 e Ω_1, podendo ocorrer devido a custos elevados para a obtenção da resposta, ou simplesmente devido a ausência de uma logística confiável. Classificadores, cuja aplicação sejam mais versáteis são utilizadas nesse processo de categorização.

Seja D um classificador que associa a cada elemento de Ω um valor d_0 ou d_1, e seja S_0 e S_1 partições de Ω, de modo que $\{D(\varepsilon) = d_0 \rightarrow \varepsilon \in S_0\}$ e $\{D(\varepsilon) = d_1 \rightarrow \varepsilon \in S_1\}$. O caso ideal, da relação entre a variável T e o classificador D, seria $S_0 = \Omega_0$ e $S_1 = \Omega_1$ indicando que D é um classificador perfeito. Porém, na grande maioria dos casos práticos, esse exercício de discriminção não é algo tão fácil de ser realizado eficientemente. Neste contexto, conceitos como o de eficiência devem ser discutos.

2.1 Medidas de eficiência

Existem medidas para avaliar o desempenho de um classificador, dentre essas, a mais simples talvez seja a **acurácia**, que consiste no número de vezes em que o mesmo classificou corretamente os objetos do estudo a serem categorizados. Normalmente é razoável pensar que acurácia alta equivale a um teste eficiente, porém existem casos em que o resultado pode enganar. Por exemplo, supondo que para um estudo epidemiológico será utilizado um classificador D para diagnosticar 100 pacientes, considerando que 5% esteja realmente doente; afirmando cegamente que todos são saudáveis, em 95% dos casos essa afirmação estará correta, resultando em alta acurácia. O que está correto para os saudáveis, mas lastimável para os doentes.

Em situações envolvendo diversos classificadores, pode-se utilizar a acurácia para compará-los. Em casos em que os mesmos obtiveram valores próximos nessa medida não necessariamente implica que ambos possuem a mesma eficiência de classificação. Por exemplo, para dois classificadores D_1 e D_2, pode ser que os erros de D_1 sejam em sua maioria na classificação em d_0 enquanto para D_2 com a classificação em d_1. Dependendo da situação um pode ser bem mais útil do que outro; diante disso, avaliar um teste apenas pela acurácia pode ser limitador, por isso, é mais proveitoso passar a analisar os resultados de maneira mais

fragmentada, a partir daí, conceitos de **sensibilidade** e **especificidade** passam a ser de grande utilidade.

2.1.1 Sensibilidade e especificidade

Sensibilidade e especifidade são definidas como duas formas distintas de eficiência; a primeira referente ao número de classificados corretamente em d_1, definidos pela proporção no conjunto $\{D(\varepsilon) = d_1 \mid \varepsilon \in \Omega_1\}$, em que cada ocorrência se trata de um **Verdadeiro Positivo** (VP), a segunda é referente ao número de classificados corretamente em d_0, definidos pela proporção no conjunto $\{D(\varepsilon) = d_0 \mid \varepsilon \in \Omega_0\}$ em que cada ocorrência é denotada por **Verdadeiro Negativo** (VN). Formalizando, especificidade e sensibilidade são dadas por

$$\text{Sensibilidade} = \frac{\text{Número de Verdadeiros Positivos}}{\text{Total de Positivos}} = \frac{\#S_1 \cap \Omega_1}{\#\Omega_1},$$

$$\text{Especificidade} = \frac{\text{Número de Verdadeiros Negativos}}{\text{Total de Negativos}} = \frac{\#S_0 \cap \Omega_0}{\#\Omega_0}.$$

Alguns autores utilizam os termos **Taxa de Verdadeiro Positivo** (TVP) e **Taxa de Verdadeiro Negativo** (TVN) para se referir a sensibilidade e especificidade respectivamente. No decorrer do trabalho serão utilizados tanto os termos TVP e TVN quanto sensibilidade e especificidade.

Um classificador dificilmente conseguirá classificar com perfeição todos os elementos do estudo, sendo possível a ocorrência de erros de classificação, nesse caso existem dois tipos de erro: $\{D(\varepsilon) = d_1 \mid \varepsilon \in \Omega_0\}$ será denotado por **Falso Positivo** (FP) e $\{D(\varepsilon) = d_0 \mid \varepsilon \in \Omega_1\}$, que será denotado por **Falso Negativo** (FN). Com isso, é possível definir **Taxa de Falso Positivo** (TFP) e **Taxa de Falso Negativo** (TFN) como dois termos antagônicos à sensibilidade e especificidade, dados por

$$\text{TFP} = \frac{\text{Número de Falsos Positivos}}{\text{Total de Negativos}} = \frac{\#S_1 \cap \Omega_0}{\#\Omega_0},$$

$$\text{TFN} = \frac{\text{Número de Falsos Negativos}}{\text{Total de Positivos}} = \frac{\#S_0 \cap \Omega_1}{\#\Omega_1}.$$

Em situações práticas, o acesso a todas as observações pode ser inviável ou praticamente impossível devido alguns fatores citados em Cochran (1977) como: necessidade de coleta em um curto espaço de tempo, um alto custo associado a coleta de cada unidade populacional que inviabiliza a realização de um censo por exemplo, dificuldades associadas a supervisão da coleta

que se torna mais complicada quando maior for a população. Como alternativa uma amostra probabilística de Ω é coletada, de modo que $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_a\} \subset \Omega$. Por amostra probabilística entende-se como procedimentos onde cada possível amostra distinta $(A_1, A_2, \ldots, A_v)$ possui probabilidade conhecida de ocorrer (COCHRAN, 1977).

As frequências apresentadas para TVP, TVN, TFP e TFN, podem ser utilizadas como estimativas para as verdadeiras taxas populacionais, em que os valores são dados por: $\mathbb{P}[D(\varepsilon) = d_1 \mid \varepsilon \in \Omega_1]$, $\mathbb{P}[D(\varepsilon) = d_0 \mid \varepsilon \in \Omega_0]$, $\mathbb{P}[D(\varepsilon) = d_1 \mid \varepsilon \in \Omega_0]$ e $\mathbb{P}[D(\varepsilon) = d_0 \mid \varepsilon \in \Omega_1]$.

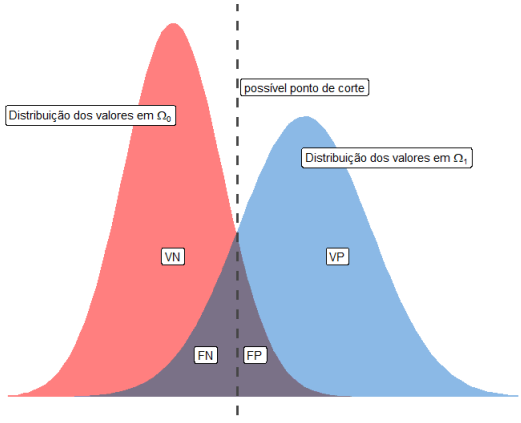
2.1.2 Ponto de Corte

O processo para definição de D pode ser obtido por intermédio de uma variável numérica W definida nos reais, com domínio em $\Omega, W : \Omega \to \mathbb{R}, W(\varepsilon) \in \mathbb{R}, \forall \varepsilon \in \Omega$; um **ponto de corte** c é definido como um critério de classificação tal que

$$\begin{cases} D(\varepsilon) = d_0, \text{ se } W(\varepsilon) < c, \\ D(\varepsilon) = d_1, \text{ se } W(\varepsilon) \geq c. \end{cases}$$

O valor atribuído a este ponto é arbitrário, porém dependendo da escolha pode haver várias combinações de decisões corretas e incorretas. Havendo uma predominância de valores positivos para valores mais altos de W, assumindo um ponto de corte muito alto, o número de verdadeiros positivos e falsos positivos serão reduzidos, por outro lado, aumentarão os casos de verdadeiros negativos e falsos negativos. É possível associar-se a esse dilema uma balança em que o objetivo é encontrar um ponto ótimo e de equilíbrio entre sensibilidade e especificidade. Essencialmente um ponto de corte é tido como um critério de divisão para as distribuições de valores de W nos elementos de Ω_0 e Ω_1, isso pode ser visualizado por meio da Figura 1

Figura 1 – Distribuições de valores hipotéticos para elementos Ω_0 e Ω_1.



No exemplo ilustrativo da Figura 1 é perceptível que para valores altos em Ω_1 e baixos em Ω_0, o processo de classificação ocorre sem muitas complicações, porém, existe uma frequência de elementos de Ω_0 e Ω_1 com valores similares para W, sendo representado pela área de encontro das duas densidades, sendo assim dado um ponto de corte qualquer, os elementos de Ω_1 que obtiveram valores mais baixos para W serão classificados como d_0 que é um erro de classificação, de forma análoga, elementos de Ω_0 que tem valores mais altos de W serão classificados como d_1 que também é um erro. Além disso é possível entender que é impossível encontrar um ponto de corte que sirva como um critério de classificação perfeito.

2.2 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é um ferramental gráfico originalmente desenvolvido durante a segunda guerra mundial no contexto de detecção de sinais eletrônicos e problemas com radares. Basicamente o objetivo era quantificar a habilidade dos operadores de radares, chamados de *receiver operators*, em distinguir se algo que foi captado pelo radar era um avião inimigo (sinal) ou algum outro objeto sem relevância (ruído). Posteriormente, a teoria de detecção de sinais foi desenvolvida por volta das décadas de 50 e 60 (GREEN; SWETS, 1966; EGAN, 1975). O potencial de utilidade da curva ROC em diagnósticos médicos foi reconhecido por volta de 1960 (LUSTED, 1960). Na década de 80 sua popularidade especialmente em radiologia cresceu após a publicação de Swets e Pickett (1982). A utilização da curva ROC se disseminou em várias outras áreas de conhecimento, como economia, para avaliação de desigualdade de renda, validação de modelos de risco de crédito; previsão de tempo para se avaliar a qualidade de predições de eventos raros; em aprendizagem de máquina e mineração de dados como um ferramental para avaliação de classificadores.

Em Bamber (1975) o problema de detecção de sinais é descrito como um experimento em que existem dois possíveis eventos: sinal e ruído. Um observador recebe uma tarefa de diferenciar ambos, assumindo que o experimento conta com j ensaios, cada observador precisa identificar se no ensaio ocorre um sinal ou ruído baseado no seu grau de certeza, além de um critério de tolerância. Se em um determinado ensaio, seu grau de certeza superar seu critério de tolerância ele responderá "sim" para sinal e caso contrário "não". Seja $\mathbb{P}(\text{sim} \mid \text{sinal})$ e $\mathbb{P}(\text{sim} \mid \text{ruído})$ as probabilidades de se dizer "sim" em ensaios com sinal e ruído respectivamente. Se o critério de tolerância for reduzido, $\mathbb{P}(\text{sim} \mid \text{sinal})$ e $\mathbb{P}(\text{sim} \mid \text{ruído})$ se tornarão maiores. Com o aumento do critério de tolerância as probabilidades serão cada vez menores. Seja I_r e I_s

variáveis aleatórias que representam o grau de certeza do observador nos dois possíveis eventos. Seja cr o critério de tolerância adotado, então

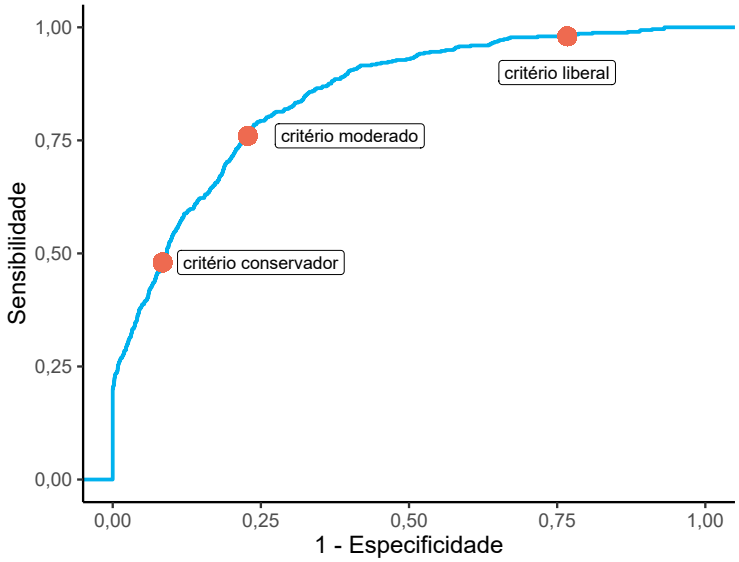
$$\begin{aligned} \mathbb{P}(\text{sim} \mid \text{sinal}) &= \mathbb{P}(I_s \geq cr), \\ \mathbb{P}(\text{sim} \mid \text{ruído}) &= \mathbb{P}(I_r \geq cr), \end{aligned}$$

realizando um paralelo com o conceitos de medidas de eficiência apresentados, é fácil ver que $\mathbb{P}(\text{sim} \mid \text{sinal})$ e $\mathbb{P}(\text{sim} \mid \text{ruído})$ são análogos as definições de sensibilidade e TFP, e cr ao ponto de corte c.

Diferentes pontos de corte irão resultar em diversos valores de sensibilidade e TFP ($1-$ especificidade). A curva ROC ilustra o comportamento de ambos a medida que o valor de c aumenta, a escolha de um ponto de corte conservador resultará em baixa sensibilidade e alta especificidade, ou seja, o classificador será mais relutante em classificar um dado elemento como d_1, o critério moderado servirá como um ponto de equilíbrio entre ambos e o critério liberal ocasionará em alta sensibilidade e baixa especificidade, nesse caso o classificador terá mais liberdade para classificar os elementos como d_1 porém isso irá resultador muitos falsos positivos.

Em algumas aplicações o custo de falsos positivos é menor do que falsos negativos ou vice versa, a curva ROC auxilia na escolha do ponto de corte que irá diferenciar de uma situação para outra.

Figura 2 – Curva ROC com três possíveis pontos de corte



Na Figura 2 percebe-se que a curva é representada no plano cujos eixos vertical e horizontal são sensibilidade e TFP respectivamente, ambos partem do ponto (0,0), em que o

delimitador W sempre assume valores menores que c e por consequência D será sempre igual a d_0, dessa maneira todas as unidades experimentais são alocados em S_0 de modo que $S_0 \cap \Omega_0 = \Omega_0$ e $S_1 = \emptyset$ tal que $S_1 \cap \Omega_1 = \emptyset$, ou seja, todos as elementos de Ω_0 estão alocadas corretamente, em contra partida, nenhum elemento de Ω_1 foi selecionado de maneira correta; por fim para ponto (1,1), o delimitador W é sempre maior do que c e D igual a d_1 em todos os casos, resultando em $S_1 \cap \Omega_1 = \Omega_1$ e $S_0 = \emptyset$, $S_0 \cap \Omega_0 = \emptyset$. Como apresentado em Dodd e Pepe (2003), sensibilidade em um dado ponto de corte c e TVP(c), podem ser definidas como $\mathbb{P}[W(\varepsilon) > c \mid \varepsilon \in \Omega_1]$ e TFP(c) como $\mathbb{P}[W(\varepsilon) > c \mid \varepsilon \in \Omega_0]$, portanto, a curva ROC é determinada como o conjunto de pontos verdadeiros e falsos positivos associados à dicotomização de W para diferentes pontos de corte c (PEPE, 2003), sendo assim, a curva ROC é dada por

$$\text{ROC}(c) = \{(\text{TFP}(c), \text{TVP}(c)), c \in (-\infty, \infty)\},$$

em que para um extremo, a medida que o valor de c aumenta

$$\lim_{c \to \infty} \text{TVP}(c) = 0 \quad \text{e} \quad \lim_{c \to \infty} \text{TFP}(c) = 0,$$

por outro, a medida que c diminui

$$\lim_{c \to -\infty} \text{TVP}(c) = 1 \quad \text{e} \quad \lim_{c \to -\infty} \text{TFP}(c) = 1.$$

A curva ROC também pode ser escrita como

$$\text{ROC}(t) = \{(t, \text{ROC}(t)), t \in (0,1)\},$$

em que $t = \text{TFP}(c)$ e $\text{ROC}(t) = \text{TVP}(c)$.

2.2.1 Propriedades matemáticas da curva ROC

Dado uma amostra, em que já se sabe se os elementos são positivos ou negativos, para uma análise dos valores obtidos na aplicação da variável W, é possível definir dois conjuntos de valores:

$$\begin{cases} X = W(\varepsilon_i), & \text{em que } \varepsilon_i \text{ é um caso negativo;} \\ Y = W(\varepsilon_i), & \text{em que } \varepsilon_j \text{ é um caso positivo.} \end{cases}$$

Esta amostra pode ser utilizada como referência para a classificação de outras observações amostrais, passíveis de classificação em positivo ou negativo. Seja $X = (x_1, x_2, \ldots, x_n)$ e $Y = (y_1, y_2, \ldots, y_m)$, dado que o processo amostral é probabilístico, pode-se supor que x_i são

valores observados de uma variável aleatória contínua X com função densidade de probabilidade $f_X(\cdot)$ e y_j são valores observados de uma variável aleatória contínua Y com função densidade de probabilidade $f_Y(\cdot)$, em que uma variável aleatória é definida por: seja $A = \{a_1, \ldots, a_s\}$ um espaço amostral em que uma função de probabilidade é definida, se Z é um função de valores reais definida em S, tal que Z associa pontos de S a valores reais, então Z é uma variável aleatória unidimensional (MOOD, 1950).

Se Z é uma variável aleatória contínua, então existe uma função não negativa $f_Z(z)$, definida para qualquer real $z \in (-\infty, \infty)$, tal que para um evento A qualquer

$$P[Z \in A] = \int_A f_Z(z)dz.$$

A função $f_Z(\cdot)$ é denominada função densidade de probabilidade da variável aleatória Z (MOOD, 1950). Além disso, outras definições que serão utilizadas no decorrer do trabalho são

$$F_Z(a) = P[Z \leq a] = \int_{-\infty}^{a} f_Z(z)dz;$$

$$S_Z(a) = P[Z \geq a] = \int_{a}^{\infty} f_Z(z)dz = 1 - F_Z(a),$$

sendo $F_Z(z)$ a função de distribuição acumulada e $S_Z(z)$ a função de sobrevivência da variável aleatória Z.

Uma classificação ineficiente W_1 não possui capacidade discriminativa quando as variáveis aleatórias X e Y são identicamente distribuídas, portanto para qualquer valor de c, $\text{TVP}(c) = \text{TFP}(c)$. Um teste perfeito W_2 consegue classificar corretamente todos os elementos de Ω_1 e Ω_2, neste caso, para algum ponto de corte c, $\text{TVP}(c) = 1$ e $\text{TFP}(c) = 0$. Na prática a maioria dos testes tem desempenho entre o teste perfeito e o não discriminativo.

Como definido em Pepe (2003), é possível estabelecer algumas propriedades formais para a curva ROC:

Propriedade 1: Seja $(\text{TFP}(c), \text{TVP}(c))$ um ponto na curva ROC para um classificador W qualquer, seja $\mathscr{T} = h(W)$ e $p = h(c)$, em que $h(.)$ é uma função estritamente crescente, então $\mathbb{P}[\mathscr{T}(\varepsilon) \geq p \mid \varepsilon \in \Omega_0] = \mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_0]$ e $\mathbb{P}[\mathscr{T}(\varepsilon) \geq p \mid \varepsilon \in \Omega_1] = \mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_1]$. Portanto, a curva ROC é invariante a transformações estritamente crescentes.

Propriedade 2: Seja S_x e S_Y funções de sobrevivência das variáveis aleatórias X e Y, tal que $S_X(c) = \mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_0]$ e $S_Y(c) = \mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_1]$. Sendo assim a curva

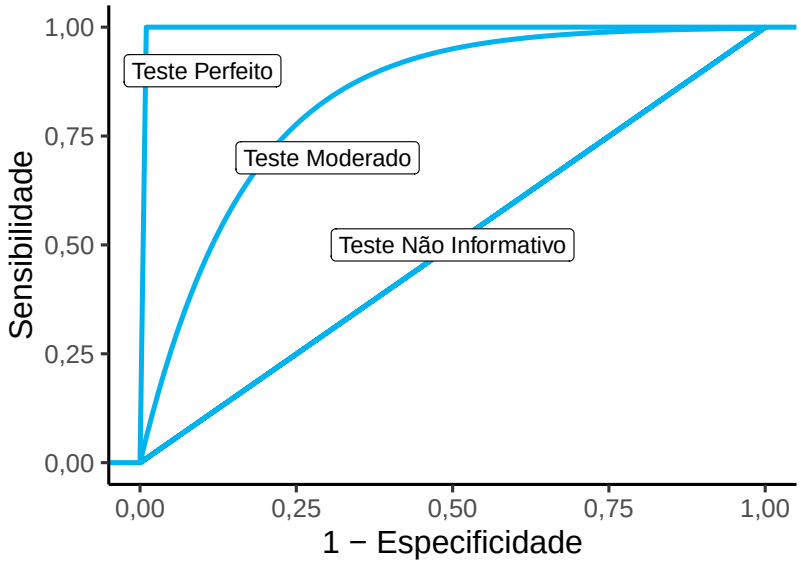
ROC pode ser representada da seguinte forma

$$\text{ROC}(c) = S_Y\left(S_X^{-1}(c)\right) \tag{2.2.1}$$

Essa definição pode ser verificada simplesmente assumindo que $c = S_X^{-1}(t)$, sendo c o ponto de corte que gera um valor de $\text{TFP}(c) = t$, de modo que $\mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_0] = t$. A TVP correspondente é $\mathbb{P}[W(\varepsilon) \geq c \mid \varepsilon \in \Omega_1] = S_Y(c)$, daí a Equação 2.2.1 pode ser facilmente verificada. Portanto, a curva ROC é formada por um par de pontos dependentes de c.

A curva ROC é uma função monotônica crescente, o seu comportamento irá variar dependendo da eficiência do classificador a ser analisado. Na Figura 3 é possível ver o comparativo das curvas ROC de três classificadores diferentes.

Figura 3 – Curvas ROC para três diferentes testes.



2.2.2 Relações de dominância

É possível estabelecer algumas relações de dominância para comparar diferentes curvas ROC, generalizando as definições apresentadas em Liu *et al.* (2005), sendo ROC_1 e ROC_2 curvas associadas a dois classificadores W_1 e W_2, é possível estabelecer as seguintes relações de dominância:

I. ROC_1 domina ROC_2 para um valor fixo t, se $\text{ROC}_1(t) > \text{ROC}_2(t)$;

II. ROC_1 domina ROC_2 em um intervalo $[t_0, t_1]$, se $\text{ROC}_1(t) > \text{ROC}_2(t), \forall t \in [t_0, t_1]$;

III. ROC_1 domina uniformemente ROC_2, se $\text{ROC}_1(t) > \text{ROC}_2(t), \forall t \in [0, 1]$.

2.2.3 Índices de sumarização

Baseado na curva ROC existem várias medidas destinadas a auxiliar na avaliação de desempenho de um classificador. Como visto, a curva ROC apresenta valores para TVP e TFP para diferentes pontos de corte, portanto conseguir agregar as informações fornecidas pela curva em medidas resumo é algo muito prático.

2.2.3.1 Área sob a curva ROC

A área sob a curva ROC (*area under curve*, AUC), é uma medida que avalia o desempenho de um classificador para todos os possíveis pontos de corte. A expressão para a AUC é dada por

$$\text{AUC} = \int_0^1 \text{ROC}(t)\text{d}t.$$

Um teste perfeito, terá valor para AUC $= 1$, por outro lado, um teste não informativo terá AUC$= 0,5$. A área sob a curva é uma medida de avaliação mais abrangente quando comparado com as relações de dominância, adotando um teste W_1 uniformemente melhor do que W_2, de modo que

$$\text{ROC}_1(t) \geq \text{ROC}_2(t), \quad \forall t \in (0,1),$$

implica que

$$\text{AUC}_1 \geq \text{AUC}_2;$$

porém a recíproca não é verdadeira. A interpretação para a área sob a curva é de que ela trata da probabilidade de uma observação, coletada aleatoriamente de X, ser menor ou igual a uma outra observação coletada aleatoriamente de Y. Com base nisso, a expressão da AUC descrita em Bamber (1975) é dada por

$$\begin{aligned} \text{AUC} &= \int_0^1 \mathbb{P}(X \geqslant c) d\mathbb{P}(Y \geqslant c) \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X \geqslant c) f_Y(c) dc \\ &= \mathbb{P}(Y \geqslant X). \end{aligned}$$

para o caso em que X e Y são variáveis aleatórias contínuas, AUC $= \mathbb{P}(Y > X)$.

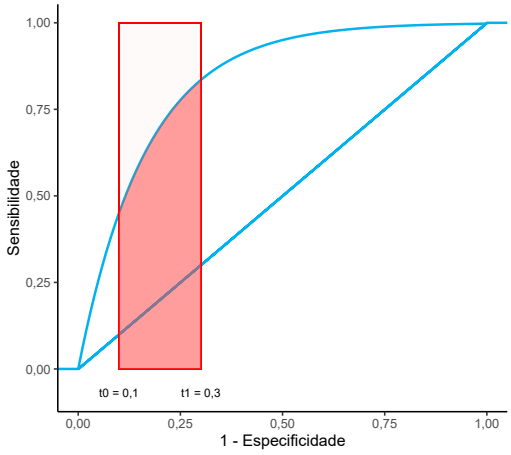
2.2.3.2 *Área Parcial sob a Curva ROC*

A AUC possui algumas limitações em determinadas aplicações, visto que ela agrega informações em regiões da curva ROC sem relevância para alguns estudos. Por exemplo, em um processo de triagem populacional, uma alta TFP gera altos custos monetários, portanto apenas a região com baixa TFP é de interesse. A área parcial sob a curva ROC (*partial* AUC, pAUC) foi abordada inicialmente em McClish (1989), Thompson e Zucchini (1989), Jiang *et al.* (1996), mas um enfoque inferencial e probabilístico mais completo foi discutido em Dodd e Pepe (2003). A pAUC definida em um intervalo (t_0, t_1) para TFP é dada por

$$\text{pAUC}(t_0, t_1) = \int_{t_0}^{t_1} \text{ROC}(t)dt, \tag{2.2.2}$$

em que $0 \leq t_0 \leq t_1 \leq 1$. A escolha do intervalo pode ser algo complicado e vai depender do prejuízo associado à TFP, que varia dependendo da situação. Portanto é necessário diálogo com o especialista da área. Uma exemplificação para a $\text{pAUC}(t_0, t_1)$ está presente na Figura 4

Figura 4 – Ilustração de uma curva ROC e sua pAUC $(t_0 = 0,1, t_1 = 0,3)$.



Alguns autores apresentaram diferentes interpretações desenvolvidas a partir da Equação 2.2.2, Dodd e Pepe (2003) apresentam uma expressão alternativa para pAUC, dada por

$$\begin{aligned}
\text{pAUC}(t_0, t_1) &= \int_{t_0}^{t_1} S_Y\left\{S_X^{-1}(t)\right\} dt = \int_{S_X^{-1}(t_1)}^{S_X^{-1}(t_0)} S_Y(y) dF_X(y) \\
&= \int_{S_X^{-1}(t_1)}^{S_X^{-1}(t_0)} S_Y(y) f_X(y) dy \\
&= \mathbb{P}\left[Y > X, X \in \left\{S_X^{-1}(t_1), S_X^{-1}(t_0)\right\}\right].
\end{aligned}$$

2.2.4 Estimação da Curva ROC

Dado que algumas propriedades da curva ROC foram definidas, existem uma série de metodologias estatísticas destinadas a realizar inferências sobre a curva ROC populacional, na literatura existem duas abordagens que aparecem com certa frequência. A primeira consiste na aplicação de métodos não paramétricos aos dados para a obtenção da curva ROC empírica. Segundo Pepe (2003) métodos empíricos são populares para problemas envolvendo classificadores contínuos. A segunda metodologia consiste na suposição de distribuições de probabilidade conhecidas para X e Y, algumas críticas surgem com relação a essa metodologia como apresentado em Pepe (2003), o principal motivo está relacionado a fortes suposições paramétricas com relação a X e Y, uma condição ocasionalmente desnecessária, visto que a curva ROC está associada a relação das distribuições de X e Y e não na forma da distribuição propriamente dita.

Algumas considerações precisam estar claras com respeito aos processo de estimação: as unidades amostrais precisam ser coletadas de maneira aleatória para garantir independência, isso será útil para definições utilizando teoria de distribuição assintótica apresentadas mais adiante. Os diferentes métodos de estimação bem como a variabilidade associada a cada um será abordado a seguir.

2.2.4.1 Estimação Empírica

O estimador empírico da curva ROC consiste da representação da mesma por meio dos dados observados, sendo assim, uma função da amostra. Portanto, para um ponto de corte c é possível estimar a TVP e TFP por

$$\begin{aligned} \widehat{\text{TVP}}(c) &= \sum_{i=1}^{m} I\left[Y_i \geq c\right] \Big/ m, \\ \widehat{\text{TFP}}(c) &= \sum_{j=1}^{n} I\left[X_j \geq c\right] \Big/ n, \end{aligned}$$

sendo $I[.]$ a função indicadora, tal que para um conjunto A

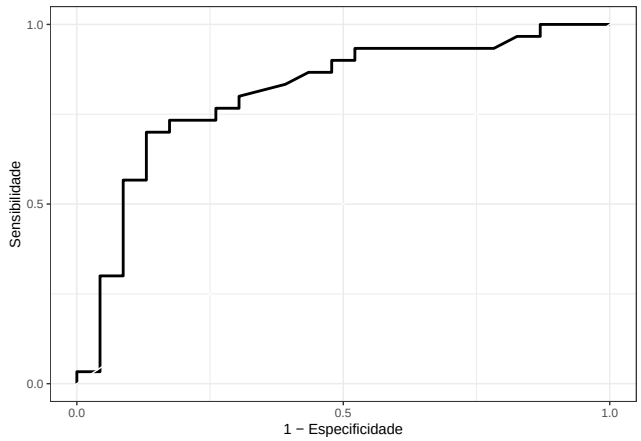
$$I_A(x) = \begin{cases} 1, & \text{se } x \in A \\ 0, & \text{se } x \notin A \end{cases}$$

Equivalente a Equação 2.2.1, a curva ROC empírica $\widehat{\text{ROC}_e}$ pode ser expressa por

$$\widehat{\text{ROC}_e}(c) = \hat{S}_Y\left(\hat{S}_X^{-1}(c)\right),$$

em que $\hat{S}_X$ e $\hat{S}_Y$ são as funções de sobrevivência empíricas das variáveis aleatórias X e Y, respectivamente. Com relação a $\widehat{\text{ROC}}_e$ sabe-se que é uma função discreta que assume valores $\{0, 1/n, 2/n, \ldots, 1\}$ no eixo horizontal referente a TFP, na prática esses ponto são unidos linearmente; quando não existem empates, ou seja, não existem valores iguais entre as observações de Ω_0, de Ω_1 ou entre Ω_0 e Ω_1, $\widehat{\text{ROC}}_e$ é um função escada crescente com saltos de tamanho $1/m$ no eixo vertical referente as observações de Ω_1 e saltos horizontais para as observações de Ω_0. Na ocorrência de empates $\widehat{\text{ROC}}_e$ apresenta características distintas dependendo do tipo de empate: (1) na ocorrência de empates entre as observações pertencentes a Ω_0 (Ω_1) ocasiona em maiores saltos no eixo horizontal (vertical), (2) empates entre os elementos de Ω_0 e Ω_1 resulta em saltos diagonais, representando um salto vertical e horizontal simultâneo. A Figura 5 apresenta a curva referente a um estudo sobre câncer de ovário abordado em Pepe (2003), em que se mediu a intensidade de um gene específico em tecidos com e sem câncer, nela é possível ver as ocorrências descritas.

Figura 5 – Curva ROC empírica para conjunto de dados sobre câncer de ovário.



2.2.4.2 Estimação da AUC empírica

A área sob a curva ROC empírica pode ser calculada utilizando a Regra dos trapézios, que consiste em dividir a curva em vários retângulos, calcular e somar suas respectivas áreas, obtendo assim um aproximação da verdadeira AUC empírica. A utilização desse método é equivalente a estatística U do teste de Mann-Whitney (MANN; WHITNEY, 1947) aplicada a duas amostras (BAMBER, 1975). Isso é bastante condizente visto que, por definição, a estatística U estima a probabilidade de que um par (x, y) coletado aleatoriamente de dois grupos distintos X e Y obedecerem a desigualdade $x < y$. Portanto o estimador não paramétrico para a AUC é dado

por

$$\widehat{\mathrm{AUC}}_e = \sum_{i=1}^{n}\sum_{j=1}^{m} \psi(x_i, y_j) \Big/ nm \tag{2.2.3}$$

sendo

$$\psi(x,y) = \begin{cases} 1, & \text{se } x < y \\ 1/2, & \text{se } x = y \\ 0, & \text{se } x > y \end{cases}$$

Basicamente o procedimento consiste em avaliar, primeiramente para X, cada uma de suas observações x_i fixa e compará-las com todas as demais de Y atribuindo uma de três possíveis classificações denotadas por C_{ij}, em que $C_{ij} = 1$ se $x_i < y_j$, que representa a classificação "correta", $C_{ij} = 0$ se $x_i > y_j$ a classificação “incorreta” e $C_{ij} = 0,5$ se $x_i = y_j$. O mesmo é feito comparando cada uma das observações y_j fixas com todas as observações de X. Um exemplo didático apresentado em Hanley e Hajian-Tilaki (1997) ajuda na compreensão desse processo; assumindo $X = (x_1, \ldots, x_9)$ e $Y = (y_1, \ldots, y_6)$ como valores de um teste para os grupos de pacientes saudáveis e doentes respectivamente.

Tabela 1 – Cálculo das classificações para o conjunto de dados com seis indivíduos doentes e nove saudáveis.

	Indivíduos do grupo doente $(m = 6)$						
Indivíduos do grupo saudável $(n = 9)$	$y_1 = 1$	$y_2 = 5$	$y_3 = 1$	$y_4 = 2$	$y_5 = 2$	$Y_6 = 5$	C_x
$x_1 = 2$	0,0	1	0,0	0,5	0,5	1	0,50
$x_2 = 1$	0,5	1	0,5	1	1	1	0,83
$x_3 = 1$	0,5	1	0,5	1	1	1	0,83
$x_4 = 1$	0,5	1	0,5	1	1	1	0,83
$x_5 = 2$	0,0	1	0,0	0,5	0,5	1	0,50
$x_6 = 1$	0,5	1	0,5	1	1	1	0,83
$x_7 = 1$	0,5	1	0,5	1	1	1	0,83
$x_8 = 1$	0,5	1	0,5	1	1	1	0,83
$x_9 = 1$	0,5	1	0,5	1	1	1	0,83
C_y	0,39	1	0,39	0,89	0,89	0,89	

Os valores presentes na Tabela 1 representam o procedimento de classificação descrito; para $x_1 = 2$ e $y_1 = 1$, $C_{11} = 0,0$ já que $x_1 > y_1$; os valores dos X-componentes denotados por C_{xi} são referentes as médias de cada uma das linhas da tabela com exceção da última e dos Y-componentes denotados por C_{yj} as médias de cada uma das colunas com exceção da última.

As medidas C_x e C_y apresentam alguns atributos interessantes; é possível utilizá-las para construir a curva ROC empírica, as médias de ambas são equivalente a área sob a curva. Posteriormente, essas medidas também serão utilizadas no estudo da variabilidade do estimador não paramétrico da AUC definido na Equação 2.2.3.

2.2.4.3 Modelagem Paramétrica

Por intermédio da Equação 2.2.1 a curva ROC pode ser estimada por meio de uma abordagem paramétrica em que se supõe distribuições de probabilidade conhecidas para X e Y. A abordagem paramétrica utilizando a distribuição normal é uma das mais utilizadas na literatura (SU; LIU, 1993; LIU *et al.*, 2005; YAN *et al.*, 2018; PEPE; THOMPSON, 2000). Supondo que $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, em que μ_x, μ_y são as médias e σ_x^2, σ_y^2 as variâncias, cada uma podendo ser estimada por seus respectivos valores amostrais. Sob tais condições a expressão para a Curva ROC é dada por

$$\text{ROC}(t) = \Phi\left(\frac{\mu_y - \mu_x}{\sigma_y} + \left(\frac{\sigma_x}{\sigma_y}\right)\Phi^{-1}(t)\right), \tag{2.2.4}$$

em que Φ denota a função de distribuição acumulada da normal padrão. Na literatura a Equação 2.2.4 é a expressão para a curva ROC binormal.

Existem algumas suposições para a curva ROC binormal decorrente da propriedade de invariância diante de transformações monotônicas estritamente crescentes, se X e Y tem distribuição de probabilidade normal e uma transformação $H_x = h(X)$ e $H_y = h(Y)$ é feita para uma função monotônica estritamente crescente, então H_x e H_y também seguem distribuição normal.

Afirmar que uma curva ROC para X e Y é binormal, é análogo a afirmação de que para uma transformação estritamente crescente h, $h(X)$ e $h(Y)$ seguem distribuição normal (PEPE, 2003). Sendo assim, é possível aplicar funções que ajudem a transformar a distribuição dos dados para um quadro mais próximo da normalidade.

A estimação para a curva definida na Equação 2.2.4 pode ser feito por intermédio dos valores amostrais da média e variância das distribuições de X e Y, denotadas por $\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_x$ e $\hat{\sigma}_y$

$$\widehat{\text{ROC}}(t) = \Phi\left(\frac{\hat{\mu}_y - \hat{\mu}_x}{\hat{\sigma}_y} + \left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\Phi^{-1}(t)\right)\right). \tag{2.2.5}$$

A AUC paramétrica é dada por

$$\text{AUC}_p = \Phi\left(\frac{\mu_y - \mu_x}{\sqrt{\sigma_x + \sigma_y}}\right),$$

que pode ser estimada por

$$\widehat{\text{AUC}}_p = \Phi\left(\frac{\hat{\mu}_y - \hat{\mu}_x}{\sqrt{\hat{\sigma}_x + \hat{\sigma}_y}}\right).$$

2.2.5 Variabilidade Amostral da Curva ROC

O estudo da variabilidade amostral para medidas associadas a curva ROC como sensibilidade, especificidade, AUC, pAUC e ponto de corte, é de grande importância para auxiliar nas interpretações dos resultados e avaliar a precisão dessas estimativas em que se observa os valores que as mesmas podem assumir caso outras amostras sejam coletadas. Em especial, para a AUC existe um conjunto de abordagens destinadas ao cálculo da variação associada ao estimador não paramétrico (HANLEY; MCNEIL, 1982; HANLEY; MCNEIL, 1983), um dos métodos mais intuitivos é o definido em DeLong *et al.* (1988), o qual consiste em calcular a variância do estimador definido na Equação 2.2.1 por intermédio das quantidades C_x e C_y apresentadas na Seção 2.2.4.2. A expressão da variância é dada por

$$Var(\widehat{\text{AUC}}_e) = \frac{\text{Variância de } C_x}{n} + \frac{\text{Variância de } C_y}{m}, \tag{2.2.6}$$

em DeLong *et al.* (1988) existe um terceiro termo $\widehat{\text{AUC}}_e(1-\widehat{\text{AUC}}_e)/nm$, sendo ignorado pelo autor por ser insignificante a medida que n e m aumentam. O fato da $\widehat{\text{AUC}}_e$ ser equivalente à estatística U permite usufruir das propriedades da mesma, dentre estas, a propriedade de normalidade assintótica, Mann e Whitney (1947) mostraram que se X e Y são variáveis aleatórias contínuas e identicamente distribuídas, a medida que n e m cresce, a distribuição de U se aproxima da distribuição normal. Posteriormente, Lehmann (1951) provou para uma classe de estatísticas, que continha U, que independente de X e Y serem identicamente distribuídas, a distribuição de todas as estatísticas desta classe é assintoticamente normal. Por meio dessas propriedades é possível definir um intervalo de confiança para a AUC, dado por

$$\widehat{\text{AUC}}_e \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{Var}(\widehat{\text{AUC}}_e)}, \tag{2.2.7}$$

em que Φ^{-1} denota a função de quantis da normal padrão, $\widehat{Var}$ a variância amostral e α o nível de significância.

Algumas considerações com respeito a intervalos de confiança na forma da Equação 2.2.7 foram feitas por Bamber (1975), que chama atenção para o fato da violação da suposição de normalidade, visto que a área sob a curva somente assume valores no intervalo $[0,1]$, ao contrário da distribuição normal que possui suporte nos reais, diante desta problemática, quando

a estimativa pontual para a AUC estiver próxima dos limites zero ou um, a suposição de normalidade será violada. Para valores distantes dos extremos, essa aproximação tenderá a ser adequada. Outro fator a se avaliar também é o tamanho da amostra, visto que quando maior for, menor será a variância definida na Equação 2.2.6 que resultará em intervalos mais estreitos.

Frequentemente é desejável saber se a área sob a curva de dois classificadores (AUC_1 e AUC_2) distintos são iguais, isso é equivalente a testar se a diferença entre ambas é igual a zero. Com isso elabora-se um teste de hipótese cujas alternativas são definidas na Equação 2.2.8. A metodologia desenvolvida por DeLong *et al.* (1988) é voltada a comparação da AUC, no âmbito não paramétrico, de curvas ROC correlacionadas, que ocorre, por exemplo, quando diferentes medidas são retiradas de uma mesma observação.

$$\begin{cases} H_0: \quad \text{AUC}_1 = \text{AUC}_2, \\ H_1: \quad \text{AUC}_1 \neq \text{AUC}_2. \end{cases} \tag{2.2.8}$$

Para a realização de um teste de hipóteses se faz necessário a definição de uma estatística de teste, nesse caso, essa estatística é dada em função da variação da diferença entre a $\widehat{\text{AUC}}_{e1}$ e $\widehat{\text{AUC}}_{e2}$, que são os estimadores não paramétricos para AUC_1 e AUC_2 respectivamente. A expressão da variância de $\widehat{\text{AUC}}_{e1} - \widehat{\text{AUC}}_{e2}$ é dada por

$$Var(\widehat{\text{AUC}}_{e1} - \widehat{\text{AUC}}_{e2}) = Var(\widehat{\text{AUC}}_{e1}) + Var(\widehat{\text{AUC}}_{e2}) - 2Cov(\widehat{\text{AUC}}_{e1}, \widehat{\text{AUC}}_{e2}),$$

em que a expressão para covariância em 2.2.9 é dada por

$$\begin{aligned} Cov(\widehat{\text{AUC}}_1, \widehat{\text{AUC}}_2) &= \frac{\text{Covariância entre os pares } C_x}{n} \\ &+ \frac{\text{Covariância para entre os } C_y}{m}. \end{aligned} \tag{2.2.9}$$

Sen (1960) apresenta um método destinado a obtenção de estimativas consistentes para os elementos da matriz de variância e covariância de uma vetor de estatísticas U . Por meio disso, DeLong *et al.* (1988) desenvolve o teste de comparação de duas ou mais AUC's com base no componente-X e no componente-Y previamente apresentados. A covariância definida na Equação 2.2.9 pode ser estimada por

$$\begin{aligned} \widehat{Cov}(\widehat{\text{AUC}}_1, \widehat{\text{AUC}}_2) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(C^1_{x_i} - \bar{C}^1_x\right)\left(C^2_{x_i} - \bar{C^2}_x\right)\right] \\ + \frac{1}{m}\left[\frac{1}{m-1}\sum_{j=1}^{m}\left(C^1_{y_j} - \bar{C}^1_y\right)\left(C^2_{y_j} - \bar{C}^2_y\right)\right] \end{aligned}$$

sendo C^1_{xi}, C^1_{yi} os X-componentes e Y-componentes para o primeiro classificador, C^2_{xi}, C^2_{yi} os componentes para o segundo classificador, $\bar{C}^1_x, \bar{C}^2_x, \bar{C}^1_y$ e $\bar{C}^2_y$ as médias dos componentes.

Em geral não existe um teste exato para a Equação 2.2.8, diante disso, a estatística para o teste assintótico desenvolvida em DeLong *et al.* (1988) para o comparativo de duas AUC's é dado por

$$z = \frac{\widehat{\mathrm{AUC}_{e1}} - \widehat{\mathrm{AUC}_{e2}} - (\mathrm{AUC}_1 - \mathrm{AUC}_2)}{\sqrt{\widehat{Var}(\widehat{\mathrm{AUC}_{e1}} - \widehat{\mathrm{AUC}_{e2}})}}, \tag{2.2.10}$$

sob hipótese nula z segue uma distribuição normal padrão.

Outra forma presente na literatura para comparação de AUC's é por meio do método de *bootstrap*. Em que se define uma quantidade Z originalmente proposta por Hanley e McNeil (1983), dada por

$$Z = \frac{\theta_1 - \theta_2}{dp(\theta_1 - \theta_2)},$$

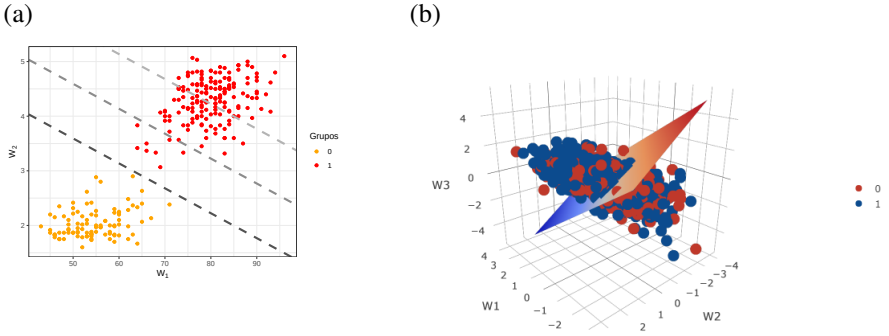
sendo θ_1 e θ_2 as duas AUC's e dp o desvio padrão. O $dp(\theta_1 - \theta_2)$ é computado com base em R réplicas *bootstrap*. Para cada réplica r, se faz uma amostragem aleatória respeitando a quantidade de indivíduos de ambos os grupos de classificação Ω_0 e Ω_1. Uma curva ROC para cada classificador é construída, calcula-se suas respectivas AUC's, $\theta_{1,r}$ e $\theta_{2,r}$, bem como sua diferença $D_r = \theta_{1,r} - \theta_{2,r}$. Por fim, $dp(D)$ é computado. Como Z segue uma distribuição normal padrão, os níveis descritivos são calculados de acordo com a hipótese nula de interesse. O método de *bootstrap* também pode ser utilizado para construção de intervalos de confiança não somente da AUC, mas também para outras medidas como pAUC, ponto de corte e AUC paramétrica (ROBIN *et al.*, 2011).

3 COMBINAÇÃO LINEAR DE CLASSIFICADORES

Em um cenário multidimensional envolvendo classificadores com capacidades discriminantes distintas, ou seja, métricas associadas a características específicas dos indivíduos de uma população, pode ser proveitoso utilizar estes classificadores simultaneamente por intermédio de um método de combinação a fim de elaborar um novo de maior eficiência. Além disso, dados multidimensionais podem ser difíceis de se analisar; a redução dessa dimensionalidade pode ser feita por intermédio da combinação linear dos classificadores (SU; LIU, 1993).

A combinação será utilizada em conjunto com o ponto de corte para classificar os indivíduos de Ω_0 e Ω_1; no cenário 2-D em que se dispõe de dois classificadores W_1 e W_2 e se estabelece um combinação linear $\alpha_1 W_1 + \alpha_2 W_2$, a equação da reta $\alpha_1 W_1 + \alpha_2 W_2 = c$ será no caso um critério de divisão. Isso pode ser visualizado na Figura 6 (a), em que as linhas em paralelo representam a escolha de diferentes pontos de corte para a combinação linear. Em um cenário 3-D a combinação passa a projetar um plano que divide os indivíduos dos grupos de classificação como pode ser visualizado na Figura 6 (b).

Figura 6 – Combinação Linear dos classificadores W_1 e W_2 para três pontos de cortes distintos, em (a). Em (b) ilustração do comportamento da combinação linear com três classificadores W_1, W_2 e W_3 no cenário 3-D.



Na literatura existem metodologias baseadas em diferentes critérios para a elaboração de uma combinação linear. Fisher (1936) considera a combinação linear de discriminantes de modo que a razão entre a diferença de médias da combinação linear nos dois grupos e sua variância seja maximizada. Welch (1939) sugere um procedimento que minimiza a probabilidade total do erro de classificação. Também existem métodos baseados na curva ROC por ser uma ferramenta avaliativa muito utilizada em estudos de classificação, em que seus índices de sumarização apresentados nas Seção 2.2.3 são também escolhas razoáveis como critérios de maximização; Su e Liu (1993) aborda dois métodos com base na curva ROC binormal em que

um deles visa a maximização da AUC, que também é um critério adotado em Pepe e Thompson (2000) para a elaboração da combinação linear não paramétrica. Yan *et al.* (2018) discutem técnicas paramétricas e não paramétricas para a maximização da pAUC.

3.1 Combinação discriminante linear

Seja $\mathbf{X} = \boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_n$ e $\mathbf{Y} = \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ amostras aleatórias de Ω_0 e Ω_1 respectivamente, tal que $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{ik}]^\intercal, i = 1, 2, \ldots, n$, e $\boldsymbol{Y}_j = \left[Y_{j1}, \ldots, Y_{jk}\right]^\intercal, j = 1, 2, \ldots, m$ são vetores aleatórios de dimensão $k \times 1$, em que cada elemento destes é uma variável aleatória com sua própria distribuição de probabilidade marginal. Supondo agora que $\mathbf{X}$ e $\mathbf{Y}$ possuem distribuição normal multivariada tal que

$$\mathbf{X} \sim \mathscr{N}_k\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right) \quad , \quad \mathbf{Y} \sim \mathscr{N}_k\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right),$$

sendo $\boldsymbol{\mu}_x$ e $\boldsymbol{\mu}_y$ os vetores de médias e $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ as matrizes de variância e covariância de dimensões $(k \times k)$.

Su e Liu (1993) apresenta duas abordagens baseadas na relação existente entre as matrizes de variância e covariância, a primeira considera que $\mathfrak{a}^2\boldsymbol{\Sigma}_y$ é igual a $\boldsymbol{\Sigma}_x$, ou seja, elas diferem por um fator de escala desconhecido $\mathfrak{a}^2$. Utilizando essa restrição é possível elaborar uma combinação linear que resultará na curva ROC dominante em relação as demais combinações lineares; a segunda abordagem não estabelece restrições para $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ e adota a AUC como critério para a elaboração da combinação.

3.1.1 Curva ROC dominante para Σ_x e Σ_y proporcionais

Seja $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]^\intercal$ o vetor de coeficientes associado à combinação linear. É possível definir duas variáveis unidimensionais $U = \boldsymbol{\alpha}^\intercal\mathbf{X}$ e $V = \boldsymbol{\alpha}^\intercal\mathbf{Y}$, tal que

$$U = \boldsymbol{\alpha}^\intercal\mathbf{X} \sim \mathscr{N}\left(\mu_u, \sigma_u^2\right) \quad , \quad V = \boldsymbol{\alpha}^\intercal\mathbf{Y} \sim \mathscr{N}\left(\mu_v, \sigma_v^2\right),$$

em que $\mu_u = \boldsymbol{\alpha}^\intercal\boldsymbol{\mu}_x, \boldsymbol{\mu}_v = \boldsymbol{\alpha}^\intercal\boldsymbol{\mu}_y, \sigma_u^2 = \boldsymbol{\alpha}^\intercal\boldsymbol{\Sigma}_x\boldsymbol{\alpha}$ e $\sigma_v^2 = \boldsymbol{\alpha}^\intercal\boldsymbol{\Sigma}_y\boldsymbol{\alpha}$. A fim de encontrar uma combinação linear que resultará em uma curva ROC que domina todas as demais, Su e Liu (1993) define $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k] \propto (\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ como os coeficientes para a melhor combinação linear, condizente com o coeficiente discriminante linear de Fisher (FISHER, 1936). Expressões para sensibilidade e especificidade podem ser obtidas em função dessa combinação linear; seja $F_U(\cdot)$ a função de distribuição acumulada de U, em que a especificidade é definida por $F_U(c)$, e

sensibilidade definida por $H(c) = 1 - G_V(c)$, tal que $G_V(\cdot)$ é a função de distribuição acumulada de V. Portanto, para qualquer valor t fixo, existe um ponto de corte c tal que

$$F_U(c) = \Phi\left(\frac{c-\mu_u}{\sigma_u^2}\right) = t.$$

O ponto de corte pode ser definido por

$$c = \mu_u + \Phi^{-1}(t)\sigma_u,$$

por fim, a expressão para sensibilidade é dada por

$$H_a(c) = 1 - \Phi\left(\frac{c-\mu_u}{\sigma_u^2}\right) = 1 - \Phi\left(\frac{\mu_v - \mu_u + \Phi^{-1}(t)\sigma_u}{\sigma_v}\right). \tag{3.1.1}$$

Para o caso de matrizes de variância e covariância não proporcionais Anderson *et al.* (1962) apresentaram uma classe de regras para classificação de duas distribuições normal multivariada utilizando uma combinação linear. Em geral, não existe uma combinação linear que resulte em uma curva ROC uniformemente dominante, a não ser que condições especiais sejam estabelecidas, estas que na prática costumeiramente são violadas. Além das classes de regras, Anderson *et al.* (1962), apresentaram critérios para estabelecer a combinação linear, dentre eles, uma solução Bayesiana em que dada uma distribuição *a priori* para as proporções populacionais, adota-se um procedimento que minimiza a probabilidade do erro de classificação para um grupo, mantendo fixa a probabilidade para o outro. Uma alternativa abordada em Su e Liu (1993) consiste de uma combinação linear de classificadores que maximize a AUC; denotada por combinação sob o critério AUC.

3.1.2 Combinação linear sem restrições em Σ_x e Σ_y

Seja $\mathbf{X} \sim \mathscr{N}_k\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$ e $\mathbf{Y} \sim \mathscr{N}_k\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$, ou seja, não há restrições especiais para $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$. Nesse caso os coeficientes para a melhor combinação linear são $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k] \propto (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)$. A expressão da AUC para a melhor combinação linear é dada por

$$\text{AUC} = \Phi\left(\sqrt{(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)^{\mathrm{T}}(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)}\right),$$

sem perda de generalidade, é possível definir a pAUC para um intervalo de TFP com limites $[t_0, t_1]$, cuja expressão obtida por meio da Equação 3.1.1 é dada por

$$\text{pAUC}(t_0, t_1) = \int_{t_0}^{t_1} 1 - \Phi\left(\frac{\mu_v - \mu_u + \Phi^{-1}(t)\sigma_u}{\sigma_v}\right) dt. \tag{3.1.2}$$

Em muitas situações $\boldsymbol{\mu_x}, \boldsymbol{\mu_y}, \boldsymbol{\Sigma_x}$ e $\boldsymbol{\Sigma_y}$ são desconhecidos, sendo possível estimá-los por meio de uma amostra representativa.

3.1.3 Processo de estimação

Seja S a soma de quadrados dada por

$$\boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_x)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_x)^\intercal + \sum_{j=1}^{m} (\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_y)(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_y)^\intercal,$$

em que $\hat{\boldsymbol{\mu}}_x$ e $\hat{\boldsymbol{\mu}}_y$ são os vetores de médias amostrais. Para para o caso em que $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}$, Su e Liu (1993) define um estimador não viciado para $\boldsymbol{\Sigma}^{-1}$ dado por $\hat{\boldsymbol{S}} = (m+n-k-3)\boldsymbol{S}^{-1}$, além disso, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_x)$ é um estimador não viciado para $\boldsymbol{\alpha}$.

Quando não vale supor que $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ são proporcionais, ambas podem ser estimadas utilizando $\boldsymbol{S}_x = \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_x)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_x)^\intercal$ e $\boldsymbol{S}_y = \sum_{j=1}^{m} (\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_y)(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_y)^\intercal$, supondo que

$$\begin{aligned} \boldsymbol{S}_x &\sim \text{Wishart}(n-1, k, \boldsymbol{\Sigma}_x), \\ \boldsymbol{S}_y &\sim \text{Wishart}(m-1, k, \boldsymbol{\Sigma}_y), \end{aligned}$$

sendo $\boldsymbol{S}_x/(n-1)$ e $\boldsymbol{S}_y/(m-1)$ estimadores de máxima verossimilhança para $\boldsymbol{\Sigma}_x$ e $\boldsymbol{\Sigma}_y$ com erro quadrático médio na ordem de $1/n$ e $1/m$, respectivamente. Su e Liu (1993) define um estimador consistente para $\boldsymbol{\alpha} = (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)$ dado por

$$\hat{\boldsymbol{\alpha}} = \left(\frac{\boldsymbol{S}_x}{n-1} + \frac{\boldsymbol{S}_y}{m-1} \right)^{-1} (\hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\mu}}_x),$$

em que o erro quadrático médio associado a $\hat{\boldsymbol{\alpha}}$ é de ordem $1/\min(n,m)$.

A utilização dos métodos apresentados exige uma avaliação do comportamento dos dados e verificação da suposição de normalidade, quando esta não é válida para $\boldsymbol{X}$ e $\boldsymbol{Y}$, a transformação de Box-Cox (BOX; COX, 1964) pode ser utilizada para melhorar o ajuste. Em geral quando ocorre afastamento da distribuição normal existe certa perda de performance associada aos métodos (SU; LIU, 1993).

3.1.4 Limitações do uso da pAUC sob suposição de normalidade

Existem algumas complicações associadas a combinação linear sob o critério pAUC, primeiro que ao contrário da AUC, a pAUC não possui expressão analítica simples, além disso existe o problema de múltiplos máximos como apresentado em Hsu e Hsueh (2013). Analiticamente o processo de maximização pode ser feito por meio de uma solução que satisfaça a Equação 3.1.3 para um t fixo

$$\frac{\partial \text{pAUC}(t)}{\partial \boldsymbol{\alpha}} = 0. \tag{3.1.3}$$

Porém a matriz hessiana é de uma complexidade considerável, que implica na incerteza associada ao vetor de coeficientes encontrado, podendo o mesmo ser um ponto de máximo global ou um ponto de mínimo local (HSU; HSUEH, 2013).

Uma expressão alternativa para um caso envolvendo um único classificador foi proposta por Thompson e Zucchini (1989). Assumindo que $X \sim \mathcal{N}(0,1)$ e $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$, a pAUC$(0,t)$ é dada por

$$\text{pAUC}(t) = \text{F}_{BVN}\left(\frac{\mu}{\sqrt{1+\sigma^2}}, \Phi^{-1}(t); -\frac{1}{\sqrt{1+\sigma^2}}\right),$$

em que $\text{F}_{\text{BVN}}(\mathfrak{v}_1, \mathfrak{v}_2; \rho) = \mathbb{P}(\mathfrak{V}_1 < \mathfrak{v}_1, \mathfrak{V}_2 < \mathfrak{v}_2)$ é a função de distribuição de variáveis aleatórias $\mathfrak{V}_1$ e $\mathfrak{V}_2$ que conjuntamente seguem normal multivariada com correlação ρ. Generalizando para o caso de múltiplos classificadores, Yan *et al.* (2018) apresenta a seguinte expressão

$$\text{pAUC}(t) = \text{F}_{\text{BVN}}\left(\frac{\boldsymbol{\alpha}^{\text{T}}(\boldsymbol{\mu_Y} - \boldsymbol{\mu_X})}{\sqrt{\boldsymbol{\alpha}^{\text{T}}(\boldsymbol{\Sigma_X} + \boldsymbol{\Sigma_Y})\boldsymbol{\alpha}}}, \Phi^{-1}(t); -\frac{\sqrt{\boldsymbol{\alpha}^{\text{T}}\boldsymbol{\Sigma_X}\boldsymbol{\alpha}}}{\sqrt{\boldsymbol{\alpha}^{\text{T}}(\boldsymbol{\Sigma_X} + \boldsymbol{\Sigma_Y})\boldsymbol{\alpha}}}\right).$$

3.1.4.1 Problema dos múltiplos máximos

O cálculo da combinação ótima sob o critério da pAUC pode envolver problema de múltiplos máximos, em que existem dois ou mais vetores de coeficientes distintos que levam a mesma curva ROC, e por consequência, a mesma pAUC. Nesse caso se diz que os vetores são equivalentes. Além disso, existe o problema de máximos locais provocados por diferentes vetores de coeficientes que não são essencialmente equivalentes, como é relatado em Hsu e Hsueh (2013). A fim de reduzir esses problemas Yan *et al.* (2018) apresenta alguns procedimentos que consiste em dividir a região de busca dos coeficientes em sub regiões em que se espera existir apenas um máximo local, em que o procedimento de otimização pode ser feito separadamente em cada umas dessas sub regiões, para que por fim, possam ser combinados. Esses procedimentos são apresentados a seguir

Padronizar o vetor de coeficientes: Dado um vetor de coeficientes $\boldsymbol{\alpha}$, e um escalar positivo $\mathfrak{b}$, tal que $\boldsymbol{\alpha}$ e $\mathfrak{b}\boldsymbol{\alpha}$ são equivalentes por resultarem na mesma curva ROC e pAUC; Yan *et al.* (2018) define o seguinte vetor de coeficientes padronizados

$$\boldsymbol{\alpha}_p = \frac{\text{sgn}(\alpha_{max})}{|\alpha_{max}|}\boldsymbol{\alpha}, \tag{3.1.4}$$

em que sgn(.) a função sinal e α_{max} o coeficiente com o maior valor absoluto. O maior coeficiente de $\boldsymbol{\alpha}_p$ será igual a um, e todos os demais $\in [-1,1]$. Dois vetores $\boldsymbol{\alpha}_i$ e $\boldsymbol{\alpha}_j$ serão equivalentes se suas formas padronizadas forem iguais.

Dividindo regiões viáveis: Para situações envolvendo dois classificadores $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*)$, adaptado de Pepe e Thompson (2000), o procedimento consiste (1) fixar $\alpha_1^* = 1$ e encontrar um valor para $\alpha_2^* \in [-1, 1]$ que maximize a função de interesse, (2) repetir para $\alpha_2^* = 1$ fixo e $\alpha_1^* \in [-1, 1]$, por fim (3) combinar os resultados obtidos nas buscas anteriores e encontrar os coeficientes cuja combinação seja a melhor sob o critério de maximização; essa estratégia divide a região de busca em duas sub regiões que não contém dois vetores de coeficientes equivalentes.

Para o procedimento de busca apresentado é possível estabelecer uma estratégia de troca de sinal da escore combinada, nesse caso $W = \alpha_1 W_1 + \alpha_2 W_2$, sempre que a AUC for menor que 0,5; isso permite que os coeficientes $(-1, \alpha_2)$ e $(\alpha_1, -1)$ sejam observados. Porém, como apresentado em Yan *et al.* (2018), essa estratégia pode provocar a ocorrência de múltiplos máximos, como alternativa rotinas semelhantes as (1) e (2) para $\alpha_1 = -1$ e $\alpha_2 = -1$ fixos podem ser utilizadas.

3.2 Métodos não paramétricos

Quando as suposições paramétricas são violadas, estimar a AUC e pAUC por meio das expressões apresentadas anteriormente não é aconselhável. Como alternativa, a utilização de métodos não paramétricos para o problema de maximização pode ser uma solução viável. As suposições previamente apresentadas não serão consideradas nessa seção, sendo $f_U(.)$ e $f_V(.)$ funções densidade de probabilidade desconhecidas para U e V.

3.2.1 Combinação linear não paramétrica

Por definição, a AUC representa a probabilidade de corretamente ordenar os pares (x_i, y_j) ou seja $x_i < y_j$, uma estimativa da AUC baseada nesse processo de ordenação é, na verdade, similar a estatística U do teste de Mann–Whitney. Sendo assim, o estimador não paramétrico para AUC associado à combinação linear é dado por

$$\widehat{\text{AUC}_{\text{np}}} \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} I\left[V_j \geq U_i\right] \Big/ nm, \tag{3.2.1}$$

O estimador não paramétrico da $\text{pAUC}(t_0, t_1)$ associado a combinação linear é dado por

$$\widehat{\text{pAUC}_{\text{np}}}(t_0, t_1) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} I\left[V_j > U_i \quad , \quad U_i \in (q_0, q_1)\right] \Big/ nm, \tag{3.2.2}$$

em que $q_0 = S_U^{-1}(t_0)$ e $q_1 = S_U^{-1}(t_1)$, em situações em que os quantis populacionais são desconhecidos, os empíricos deverão ser utilizados.

As Equações 3.2.1 e 3.2.2 fornecem soluções mais robustas para o problema de maximização de AUC e pAUC afim de encontrar coeficientes ótimos para a combinação linear de classificadores, porém por não serem funções contínuas, um método de busca ao invés de um por derivação é necessário para a maximização, outra complicação é de que o processo de otimização fica mais difícil a medida que o número de classificadores aumenta. Com o objetivo de diminuir o gasto computacional, Pepe e Thompson (2000) propuseram um método *stepwise*, dado por

> **Método *stepwise***: Seleciona-se os dois classificadores que juntos produzem a maior área de todos os k classificadores, com isso um novo classificador é formado, $W_1(\alpha) = Y_1 + \alpha Y_2$. Em seguida, o terceiro melhor classificador é agregado a $W_1(\alpha)$, formando sem perda de generalidade um novo classificador $W_2(\alpha,\beta) = Y_1 + \alpha Y_2 + \beta Y_3$. O procedimento é repetido até que todos os k classificadores sejam incluídos no modelo. Pepe e Thompson (2000) afirma que esse método é útil para economizar gastos computacionais, porém os coeficientes obtidos podem não ser os melhores no espaço k-dimensional.

Algumas considerações sobre a eficiência dos métodos não paramétricos foram feitas por Pepe e Thompson (2000), por meio de estudos de simulação em que se considerou dados com comportamento normal em que ambos os grupos possuíam matrizes de variância e covariância iguais, percebeu-se que o desempenho da combinação linear não paramétrica sobre o critério da AUC foi similar ao obtido utilizando a combinação linear sob suposição de normalidade proposta por Su e Liu (1993). Além disso, o processo de combinação linear, nesse cenário, trouxe maior contribuição para AUC quando os classificadores eram não correlacionados.

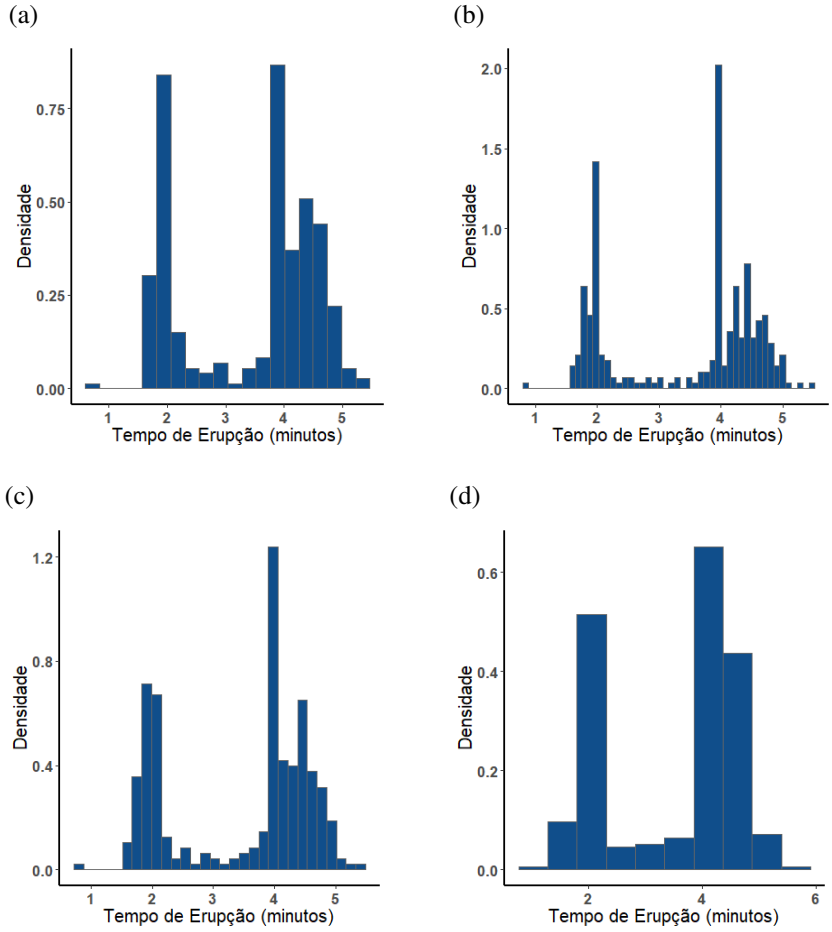
3.2.2 Combinação linear não paramétrica por kernel

A utilização de kernel pode ser útil para modelagem de um conjunto de dados sem que haja a suposição de uma distribuição de probabilidade específica para os mesmos. Esta metodologia pode ser entendida por meio do processo de construção de um histograma, que é uma forma de estimação não paramétrica da densidade de um conjunto de observações, nele d observações são agrupadas em intervalos geralmente equiespaçados de tamanho h, de modo que a densidade no ponto s é estimada por

$$\hat{f}_H(s;h) = \frac{\text{número de observações no intervalo que contém } s}{dh}.$$

Duas escolhas importantes são tomadas na construção de um histograma: (1) o valor do limite inferior do primeiro intervalo e (2) o valor h para a largura dos intervalos. Dependendo destas escolhas, o histograma terá diferentes formas, como pode ser visualizado na Figura 7.

Figura 7 – Histogramas do tempo de erupção de gêiseres no Parque Nacional de Yellowstone para diferentes valores de h.



Fonte: Azzalini e Bowman (1990).

Seguindo esse raciocínio, para uma amostra aleatória $S_1, \ldots, S_d$ de uma variável aleatória contínua e univariada S, com função densidade de probabilidade $f_S(\cdot)$, tal que

$$f_S(s) = \lim_{h \to 0} \frac{1}{2h} \mathbb{P}\left[s - h < S < s + h\right] \tag{3.2.3}$$

para qualquer valor de h, é possível estimar a função densidade de probabilidade definida na Equação 3.2.3 por meio da quantidade de observações no intervalo $(s-h, s+h)$. Portanto uma forma natural de estimar a Equação 3.2.3 para um valor pequeno de h, é dado pela expressão a seguir

$$\hat{f}_S(s) = \frac{1}{2hd} \left[\text{Número de observações no intervalo } (s-h, s+h)\right], \tag{3.2.4}$$

em que a Equação 3.2.4 é denotado na literatura como *naive estimator*, cuja expressão pode ser também definida por

$$\hat{f}_S(s) = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{h}\omega\left(\frac{s-S_i}{h}\right)$$

em que

$$\omega(s) = \begin{cases} 1/2, & \text{se } |s| < 1, \\ 0, & \text{caso contrário,} \end{cases} \tag{3.2.5}$$

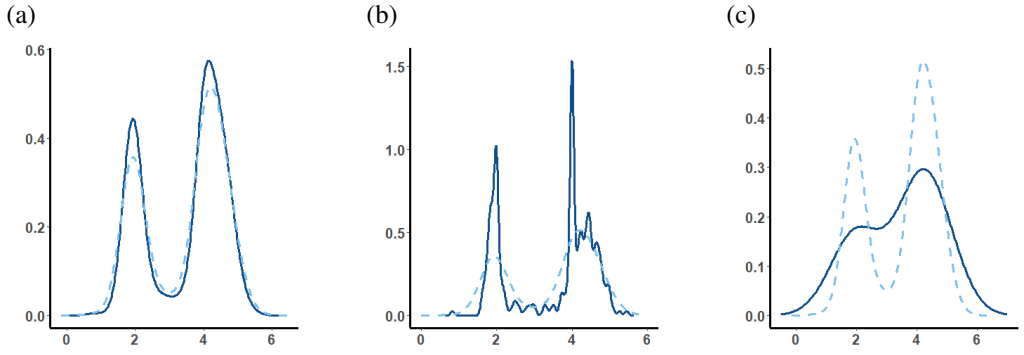
desse modo, para cada observação se constrói uma retângulo de largura $2h$ e altura $(2hd)^{-1}$.

Generalizando para o caso em que a Equação 3.2.5 é substituída por uma função K que satisfaz $\int_{\mathbb{R}} K(x)dx = 1$ conhecida como kernel, é possível definir o estimador da densidade kernel no caso univariado por

$$\hat{f}_S(s) = (dh)^{-1}\sum_{i=1}^{d} K\left\{(s-S_i)/h\right\}, \tag{3.2.6}$$

em que h é um número positivo referente a largura intervalar, também conhecido como parâmetro de suavização. Geralmente a escolha para K é de uma função densidade de probabilidade simétrica e unimodal em torno de zero, isso garante que $\hat{f}_S(x)$ definida na Equação 3.2.6 também seja uma densidade; embora existam casos em que kernels que não sejam densidades são utilizados (WAND; JONES, 1994). A escolha de K não é particularmente importante, já o valor de h tem grande impacto no processo de estimação.

Figura 8 – Gráfico das estimativas de densidade kernel para o tempo de gêiseres no Parque Nacional de Yellowstone para diferentes valores de h. A linha contínua é referente a estimativa kernel, a pontilhada à verdadeira densidade.



Fonte: Azzalini e Bowman (1990).

Na Figura 8, o gráfico (a) representa a densidade de kernel para $h = 0,25$, resultando em uma densidade próxima da real. Já (b) e (c) com valores de h de 0,05 e 0,8 respectivamente,

resultaram em estimativas inadequadas, no caso de (b) o resultado foi uma densidade pouco suave (*undersmoothed*), em (c) o contrário ocorre, que é o caso de uma estimativa da densidade super suave (*oversmoothed*).

Portanto, para o processo de análise com relação a curva ROC, é possível utilizar o estimador de kernel para estimar as funções densidade de probabilidade de U e V sem a necessidade de suposições paramétricas. Os estimadores de densidade kernel nesse caso são dados por

$$\begin{aligned} \hat{f}_U(w) &= \frac{1}{nh_U}\sum_{i=1}^{n} K\left(\frac{w-U_i}{h_U}\right), \\ \hat{f}_V(w) &= \frac{1}{mh_V}\sum_{i=1}^{m} K\left(\frac{w-V_i}{h_V}\right). \end{aligned}$$

A expressão para o estimador de kernel para AUC definida em Lloyd (1998), é dada por

$$\widehat{\mathrm{AUC}}_k = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\Phi\left(\frac{V_i-U_j}{\sqrt{h_U^2+h_V^2}}\right), \tag{3.2.7}$$

em que h_x e h_y são definidos em Zou *et al.* (1998) por

$$\begin{aligned} h_U &= 0,9\min(dp(U),\mathrm{IQR}(U)/1,34)n^{-1,5}, \\ h_V &= 0,9\min(dp(V),\mathrm{IQR}(V)/1,34)m^{-1,5}, \end{aligned}$$

sendo IQR o intervalo interquartil.

Em Yan *et al.* (2018) o estimador para pAUC por kernel é obtido utilizando $h_U = 0,9\min\left(dp(V_i),\mathrm{IQR}(V_i)\right)n^{0,2}$, com notação análoga para V. Dada a densidade estimada, é possível obter as funções de sobrevivência, definidas por

$$\begin{aligned} \hat{S}_U(w) &= \int_w^{\infty}\hat{f}_U(t)dt, \\ \hat{S}_V(w) &= \int_w^{\infty}\hat{f}_V(t)dt, \end{aligned}$$

a estimativa da curva de ROC pelo método de kernel é dada por

$$\widehat{\mathrm{ROC}}_K(t) = \hat{S}_V(\hat{S}_U^{-1}(t)),$$

por fim, o estimador de kernel para a pAUC é dado por

$$\widehat{\mathrm{pAUC}}_K(t_0) = \int_0^{t_0}\mathrm{ROC}_K(t)dt.$$

Portanto, é possível definir um método de combinação cujo objetivo seja encontrar o coeficiente $\boldsymbol{\alpha}$ que maximize a AUC definida na Equação 3.2.7, ou utilizando o método definido por Yan *et al.* (2018) para o critério pAUC. Algumas considerações para o método de kernel são destacadas em Yan *et al.* (2018): (1) tamanho da amostra, (2) tempo computacional e (3) intervalo apropriado. (1) recomenda-se uma amostra de tamanho 100 para cada um dos grupos, para gerar estimativas robustas. Em (2) o tempo computacional necessário para encontrar os coeficientes pelo abordagem de kernel é consideravelmente alto quando comparado com os demais métodos apresentados. Para (3) a escolha do intervalo h, que em certos cenários, proporciona uma estimava da área que subestima o verdadeiro valor.

4 APLICAÇÃO

Foi realizada uma análise do conjunto de dados relativo a situação cadastral dos alunos da Universidade Federal do Ceará no ano de 2018, cujo processo de ingresso dos mesmos foi feito por intermédio do SiSU dos anos de 2014 e 2015. O objetivo do estudo é analisar o fenômeno da evasão discente precoce cuja ocorrência representa um cenário catastrófico no âmbito educacional por ser muito custoso para a Instituição. Estudantes ingressos que desistem são desperdícios sociais, acadêmicos e econômicos; no setor privado significa perda de receitas, no setor público são recursos investidos sem retorno (LOBO *et al.*, 2007).

O MEC/SESU (1996) por meio da Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras classifica o fenômeno da evasão no ensino superior em:

- **Evasão de curso**: Aluno desliga-se do curso de origem sem concluí-lo (transferência interna ou aprovação no vestibular para outro curso na mesma instituição);
- **Evasão de instituição**: Aluno abandona a IES na qual está matriculado (transferência externa ou aprovação no vestibular para curso em outra instituição);
- **Evasão de sistema**: Aluno se ausenta de forma permanente ou temporária da academia (desistência).

Na análise, foi considerado evadido o aluno que teve cancelamento de matrícula no curso.

A PORTARIA NORMATIVA Nº 21, DE 5 DE NOVEMBRO DE 2012, instituída pelo MEC, que trata do Sistema de Seleção Unificada - SiSU, estabelece a possibilidade das instituições participantes designarem pesos e notas mínimas referentes às provas do ENEM. Tal procedimento visa gerar "melhores" perfis dos ingressantes nos cursos superiores. Na UFC, até o processo dos ingressantes do ano de 2018, não há pesos diferenciados entre as provas, assim como não há nota de corte. Uma possível análise sobre a eficiência da adoção de pesos diferenciados nas provas do ENEM, pode ser realizada por intermédio dos métodos apresentados nos capítulos anteriores.

No presente trabalho foram utilizadas as notas obtidas pelos alunos, nos exames de 2014 e 2015, nas disciplinas de Matemática, Ciências da Natureza, Ciências Humanas, Português e Redação; por meio destas procurou-se elaborar um classificador baseado na combinação linear das diferentes disciplinas de modo a definir uma métrica que considera pesos para cada uma das provas bem como a escolha de pontos de corte focando na implementação no sistema de seleção que pode ser adotado pela Universidade.

4.1 Apresentação dos dados

Para esse estudo foram consideradas 1569 alunos, dos quais 919 (59%) estão com matrícula ativa, 628 (40%) com matrícula cancelada e 22 (0,1%) concluíram a graduação. O conjunto de dados é apresentado na Tabela 2

Tabela 2 – Status e Notas dos alunos que realizam a prova do ENEM (2014 e 2015).

Indíce	*Status*	Por	Mat	CN	CH	Red
1	ATIVO	648,10	666,70	603,80	673,00	780,00
2	ATIVO	579,90	735,80	598,10	623,00	720,00
3	ATIVO	332,30	743,40	629,80	671,60	780,00
⋮	⋮	⋮	⋮	⋮	⋮	⋮
927	CANCELADO	596,30	722,30	607,80	745,50	660,00
928	CANCELADO	618,80	775,40	617,90	684,40	860,00
929	CANCELADO	635,30	725,70	628,30	769,60	800,00
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1567	CONCLUÍDO	620,90	784,40	644,90	665,20	660,00
1568	CONCLUÍDO	598,40	850,40	678,00	654,80	780,00
1569	CONCLUÍDO	627,60	827,30	643,80	651,20	940,00

As variáveis do estudo são

- Mat: Nota na prova de Matemática;
- Por: Nota na prova de Português;
- CN: Nota na prova de Ciências da Natureza;
- CH: Nota na prova de Ciências Humanas;
- Red: Nota na prova de Redação;
- *Status*: Indicador se o aluno evadiu ou não.

Alunos com *Status* ativo ou concluído serão considerados como igualmente pertencentes a um único grupo de classificação análogo a Ω_1, no texto será referido como grupo dos **não desistentes**, estudante com *Status* cancelado serão pertencentes ao conjunto de classificação Ω_0 ou grupo dos **desistentes**. Para essa análise foram considerados apenas os alunos dos cursos oferecidos pelo Centro de Tecnologia da UFC.

4.2 Análise descritiva

Para a análise descritiva foram calculadas primeiramente algumas medidas para cada uma das notas apresentadas na Tabela 3

Tabela 3 – Medidas Descritivas para notas do ENEM (2014 e 2015).

	Mínimo	1^o Q	Mediana	Média	3^o Q	Máximo	dp	CV
Port	332,3	596,0	622,4	620,4	646,8	758,9	42,53	0,0685
Mat	343,9	696,5	757,3	750,6	808,7	973,6	84,78	0,1129
CN	389,8	606,6	649,9	649,8	694,2	860,0	62,82	0,0966
CH	548,4	654,8	684,2	685,2	714,6	863,6	43,23	0,0631
Red	380,0	740,0	820,0	814,6	900,0	1000,0	102,62	0,1259

Para a prova de Português é possível perceber que se trata da avaliação com as menores notas, em que os valores variam de 332,3 a 758,9, as medidas de tendência central são similares com 620,4 e 622,4 para a média e mediana respectivamente, além disso pelo coeficiente de variação (0,0685) a nota de Português apresenta a segunda menor variação. A prova de Matemática apresenta maior magnitude com relação as notas quando comparada a de Português, com variação de 343,9 a 973,6, a média e mediana de 750,6 e 757,3 também indicam isso, porém sua variação é a segunda maior das provas como indicada pelo coeficiente de variação (0,1129). A prova de Ciências da Natureza é um intermediário entre Português e Matemática, na maioria das suas medidas de posição, tendência central e pelo coeficiente de variação (0,0966). As provas descritas até então trazem um comportamento similar entre si, todas apresentam amplitude considerável do mínimo para o primeiro quartil, algo que não ocorre entre o terceiro quartil e o máximo. A prova de Ciências Humanas apresenta comportamento similar com a de Português do que diz respeito a dispersão das notas, com coeficiente de variação de 0,0631 sendo o menor dentre as provas, porém a mesma apresenta magnitude de valores maior que a prova de Português, variando de 548,4 a 863,6, com mediana e média de 684,2 e 685,2 respectivamente. As notas de Redação são as que apresentam os maiores valores, variando de 380,0 a 1000,0, com mediana e média de 820,0 e 814,6, porém pelo coeficiente de variação (0,1259) a prova de Redação é a que apresenta a maior variação, além disso a amplitude entre o mínimo e primeiro quartil é considerável, similar ao que foi observado para as três primeiras provas, sugerindo assimetria dos dados.

A Tabela 4 apresenta as medidas descritivas para cada uma das provas para o grupo

dos desistentes e não desistentes. Em geral, as notas das provas dos alunos do grupo desistentes são menores que a do grupo dos não desistentes, as interpretações com base no coeficiente de variação são similares as apresentadas para a Tabela 3.

Tabela 4 – Medidas Descritivas para notas do ENEM (2014 e 2015). As linhas em branco são referentes as medidas descritivas dos alunos com matrícula cancelada, as azuis dos alunos com matrícula ativa ou concluída.

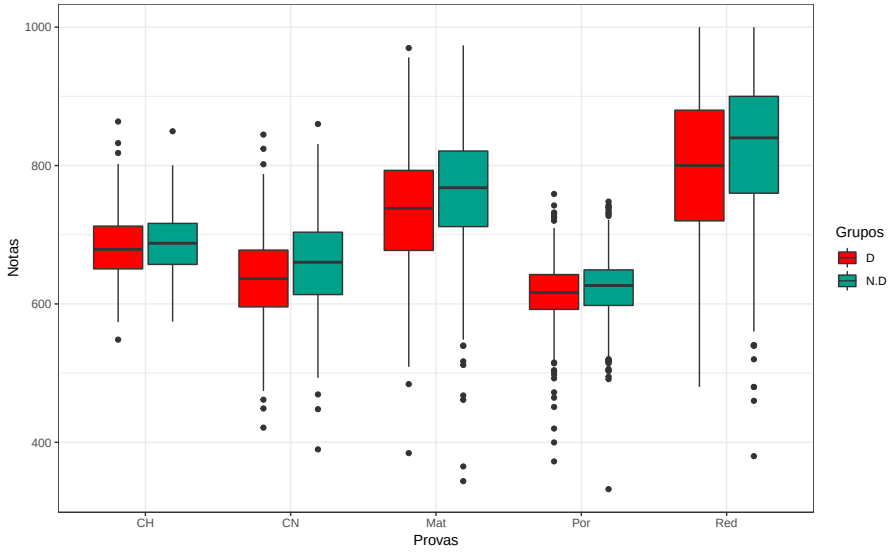
	Mínimo	1º Q	Mediana	Média	3º Q	Máximo	dp	CV
Por	372,4	592,1	616,4	615,8	642,5	758,9	43,56	0,0707
Por	332,3	597,8	626,6	623,5	649,2	747,9	41,57	0,0666
Mat	384,4	677,3	738,2	733,5	793,0	969,8	81,92	0,1116
Mat	343,9	711,7	767,9	762,0	821,0	973,6	84,78	0,1112
CN	421,2	595,7	636,5	636,7	677,8	844,7	61,21	0,0961
CN	389,8	613,5	660,2	658,5	703,7	860,0	62,39	0,0947
CH	548,4	650,6	678,8	681,8	712,5	863,6	43,69	0,0640
CH	574,4	657,1	687,7	687,4	716,4	849,5	42,79	0,0622
Red	480,0	720,0	800,0	799,8	880,0	1000,0	103,06	0,1288
Red	380,0	760,0	840,0	824,5	900,0	1000,0	101,18	0,1227

A matriz de gráficos na Figura 16 (ver Apêndice) apresenta informações sobre as notas dos alunos desistentes e não desistentes com relação a cada uma das provas, sendo composta na parte superior pelo coeficiente de correlação de Pearson, na diagonal as estimativas de densidade kernel e na parte inferior pelos gráficos de dispersão. Considerando o coeficiente de correlação percebe-se que as provas apresentam correlação moderada, sendo a maior entre Matemática e Ciências da Natureza, isso não ocorre para Redação cujo coeficiente de correlação com todas as demais é próximo a zero, indicando relação linear muito fraca. As curvas de densidade entre grupos de classificação são similares para todas as notas, sugerindo certa semelhança no perfil dos alunos desistentes e não desistentes quando se avalia as provas marginalmente isso pode ser também visualizado com auxílio da Figura 9, nela também é possível ver a diferença da magnitude das notas comentadas na análise da Tabela 3 em que a prova de Português possui em geral as menores notas e a de Redação as maiores.

A fim de definir um classificador baseado na combinação linear das provas, foram utilizados os seguintes métodos: Normal Discriminante Linear (SU; LIU, 1993) para o caso de matriz de variância e covariâncias não proporcionais, Não Paramétrica utilizando o método *stepwise* (PEPE; THOMPSON, 2000), Não paramétrico simultâneo em que todos os coeficientes são selecionados pelo algoritmo de busca de forma simultânea, Combinação Linear Não Paramé-

trica por kernel. Também foi realizado o ajuste de um modelo de regressão logística por ser uma abordagem respeitada na área da estatística e servindo também como referência.

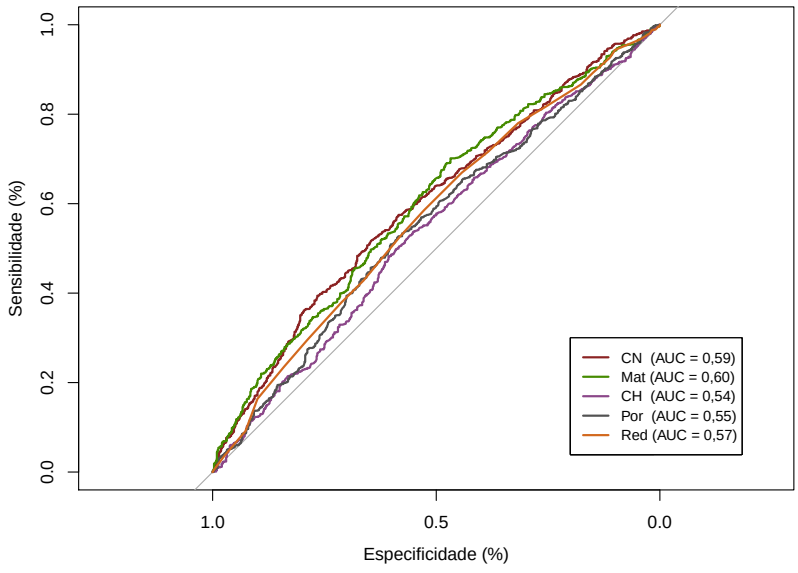
Figura 9 – BoxPlot referente as notas no ENEM para ambos os grupos de classificação. D: Alunos Desistentes, N.D: Alunos Não Desistentes.



Para essa análise todas as notas foram padronizadas. Os procedimentos computacionais foram realizados com auxilio do *software* R (R CORE TEAM, 2018).

Primeiramente, tratar cada uma das notas como classificadores é interessante para analisar como se dá o desempenho de cada umas das variáveis da combinação linear individualmente. A Figura 10 apresenta as curvas ROC para cada uma das notas.

Figura 10 – Curva ROC para classificadores individuais.



Considerando cada uma das provas como um classificador é possível perceber pela Figura 10 que todos apresentam baixa eficiência de classificação, isso pode ser compreendido com o auxílio da análise descritiva em que o comportamento das distribuições de frequência dos alunos desistentes e não desistentes para cada uma das provas são similares, isso resultará em uma quantidade considerável de erros de classificação independente da escolha do ponto de corte (relembrando a exemplificação da Figura 1). A prova de Matemática e de Ciências da Natureza apresentaram as maiores AUC's (0,60 e 0,59) dentre as disciplinas avaliadas, além disso a curva ROC de ambas domina as demais em praticamente todo o intervalo de especificidade. A prova de Ciências da Humanas ficou bem próxima da linha que indicaria que o classificador é não informativo, apresentando uma AUC (0,54) muito próxima de 0,5. Com base nos dados procurou-se elaborar uma combinação linear desses classificadores pra a elaboração de um novo classificador de maior eficiência, utilizando as diferentes metodologias apresentadas.

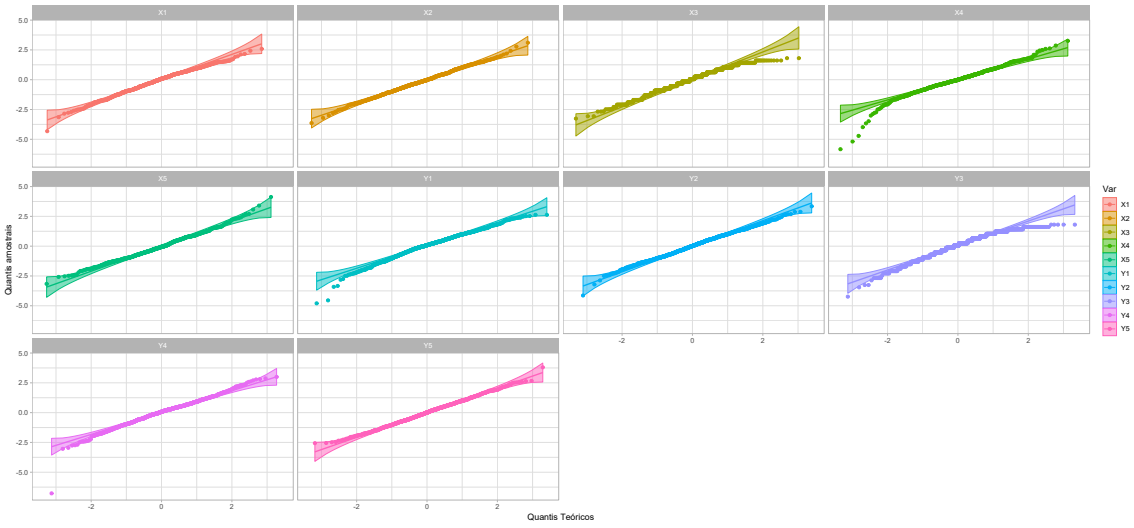
4.3 Ajuste para Modelo Normal Discriminante Linear

Para o ajuste do modelo normal discriminante linear certas suposições precisam ser avaliadas. Como visto na análise descritiva, algumas notas apresentavam assimetria à esquerda, o que sugere um afastamento da distribuição normal. Um forma de verificar a suposição de normalidade se dá pela construção de um gráfico *QQplot* com bandas de confiança construídas através de simulações de Monte Carlo propostas por Atkinson (1981), originalmente para verificação da distribuição dos resíduos do modelo de regressão normal linear, denominado de **envelope**. Os gráficos estão presentes na Figura 11, em que X_1, X_2, X_3, X_4, X_5 e Y_1, Y_2, Y_3, Y_4, Y_5 são as notas de Matemática, Ciências da Natureza, Redação, Português e Ciências Humanas para o grupo de alunos desistente e não desistente respectivamente

O comportamento esperado é de que os pontos estejam dentro das bandas de confiança próximos a reta em diagonal. Em geral visto que foram construídos intervalos de 95% de confiança, espera-se que 95% dos pontos estejam dentro dos limites. Alguns dos gráficos presentes na Figura 11 indicam afastamento da distribuição normal, principalmente para variáveis X_1, X_3, X_4, Y_1 e Y_3; a transformação de Box-Cox pode ser aplicada à essas variáveis (não transformadas) no intuito de deixá-las com comportamento mais próximo da normalidade. Porém nesse caso, as variáveis não transformadas serão mantidas visto a natureza da aplicação, em não admitir transformações mais sofisticadas para os dados. Diante desses percalços, este modelo não é o mais adequado para a análise desse conjunto de dados devido a violações em

suposições primárias, isso resulta em certa perda de eficiência como apresentado em Su e Liu (1993). Mesmo assim o ajuste será feito por motivos de comparação com os outros tipos de modelo em relação a suas eficiências de classificação.

Figura 11 – Gráficos *QQplot* para cada uma das notas avaliadas por grupo.



Portanto, assumindo que $\mathbf{X} \sim \mathcal{N}_5(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ e $\mathbf{Y} \sim \mathcal{N}_5(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, as estimativas dos vetores de médias e das matrizes de variância e covariância são dadas por

$$\hat{\boldsymbol{\mu}}_x = \begin{bmatrix} -0,20 & -0,20 & -0,14 & -0,10 & -0,07 \end{bmatrix}^\intercal$$

$$\hat{\boldsymbol{\mu}}_y = \begin{bmatrix} 0,13 & 0,14 & 0,09 & 0,07 & 0,05 \end{bmatrix}^\intercal$$

$$\hat{\boldsymbol{\Sigma}}_x = \begin{bmatrix} 0,9338 & 0,5945 & -0,0898 & 0,1483 & 0,2900 \\ 0,5945 & 0,9595 & 0,0229 & 0,3077 & 0,4267 \\ -0,0898 & 0,0229 & 1,0085 & -0,0219 & -0,0197 \\ 0,1483 & 0,3077 & -0,0219 & 1,0491 & 0,3748 \\ 0,2900 & 0,4267 & -0,0197 & 0,3748 & 1,0215 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_y = \begin{bmatrix} 0,9999 & 0,6598 & 0,0277 & 0,2521 & 0,3938 \\ 0,6598 & 0,9864 & 0,1121 & 0,3029 & 0,5305 \\ 0,0277 & 0,1121 & 0,9721 & 0,0062 & 0,0410 \\ 0,2521 & 0,3029 & 0,0062 & 0,9553 & 0,3903 \\ 0,3938 & 0,5305 & 0,0410 & 0,3903 & 0,9799 \end{bmatrix}$$

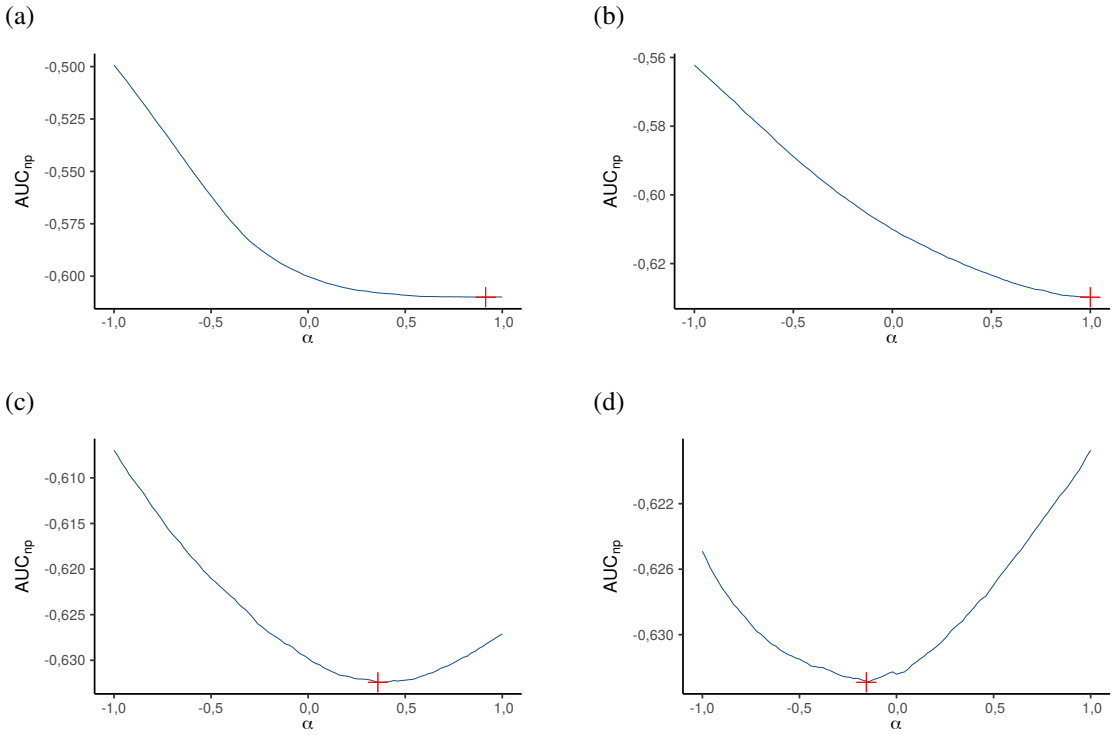
Considerando que $\hat{\boldsymbol{\Sigma}}_y$ e $\hat{\boldsymbol{\Sigma}}_x$ são não proporcionais, o vetor de coeficientes estimado é dado por

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= (\hat{\boldsymbol{\Sigma}}_x + \hat{\boldsymbol{\Sigma}}_y)^{-1}(\hat{\mu}_y - \hat{\mu}_x) = \begin{bmatrix} 0,12 & 0,1 & 0,12 & 0,05 & -0,05 \end{bmatrix} \\ &\propto \begin{bmatrix} 0,98 & 0,84 & 1,00 & 0,45 & -0,38 \end{bmatrix}\end{aligned}$$

4.4 Ajuste para Modelo Não Paramétrico com método stepwise, simultâneo e Kernel

Adotando o método *stepwise* proposto por Pepe e Thompson (2000), fixou-se para a variável com maior AUC o coeficiente $\alpha_1 = 1$ que no caso se trata da nota de Matemática, posteriormente a primeira combinação foi feita para a variável com a segunda maior AUC, sendo a nota de Ciências da Natureza, após a obtenção da escore combinada Mat$+\alpha_2 CN$, o processo foi feito para as demais provas seguindo a ordem Redação, Português e Ciências Humanas. Cada umas das etapas do processo de otimização pode ser visualizada pelos gráficos na Figura 12, que mostra o valor da $\widehat{\text{AUC}}_{\text{np}}$ a medida que o valor do respectivo coeficiente varia, tal que (a) representa $S_1 = \text{Mat} + \alpha_2 \text{CN}$, (b) $S_2 = S_1 + \alpha_2 \text{Red}$, (c) $S_3 = S_2 + \alpha_3 \text{Por}$ e (d) $S_4 = S_3 + \alpha_4 \text{CH}$.

Figura 12 – Gráficos das etapas do método *stepwise*.



Tanto o ajuste do modelo pelo método de *stepwise* quando por otimização simultânea, foram realizados utilizando a função *optim*, que conta com algumas opções com relação ao método de otimização a ser utilizado, para esta tarefa utilizou-se o método L-BFGS-B em que é

possível estabelecer limites para a busca dos coeficientes da combinação.

A combinação linear pelo método de Kernel foi feita de forma similar, em que se buscou uma combinação linear a fim de maximizar a AUC dada pela Equação 3.2.7. O resultado para os valores dos coeficientes de todos os modelos padronizados na forma apresentada pela Equação 3.1.4, estão presentes na Tabela 5.

Tabela 5 – Resultados para combinação linear das provas do ENEM de 2014 e 2015 sob as três abordagens baseadas no critério da AUC.

Modelo	Mat	CN	CH	Red	Por
Normal	0,98	0,84	-0,38	1,00	0,45
N P *stepwise*	1,00	0,91	-0,15	1,00	0,36
N P simultânea	1,00	0,57	-0,22	0,99	0,39
Kernel	1,00	0,72	-0,19	0,95	0,37

Pela Tabela 5 é possível ver que os coeficientes associados as notas de Matemática e de Redação são similares para as quatro combinações, porém para as demais notas os coeficientes variam consideravelmente. Um fato conhecido é a presença de diversos máximos locais quando a expressão 3.2.1 é utilizada para o processo de maximização, sendo uma função de indicadoras a mesma não contínua para valores de $\boldsymbol{\alpha}$, comprometendo assim o encontro de um combinação única. Muitos chutes iniciais foram testados para esse processo em vista dessa problemática. Ao menos é possível perceber certa concordância para os coeficientes das notas de CN, CH e Por, a nota de Ciências da Natureza possui coeficientes maior que a de Português, a prova de Ciências Humanas está associada a um coeficiente negativo em todos os casos, fator este que também deve ser discutido se realmente for aplicado, pois basicamente o modelo determina que quanto maior são as notas dos alunos em Ciências Humanas menor será sua escore combinada W culminando no mesmo sendo classificado como possível desistente, e mais grave, que para o sistema SiSU essa informação não poderia ser diretamente aplicada visto que alunos com notas altas em CH correriam o risco de não serem classificados. Durante o processo de maximização a nota de CH contribuiu pouco para o acréscimo na AUC que poderia ser desconsiderada, para isso um método informal apresentado em Pepe e Thompson (2000) que estabelece a construção de intervalos de confiança para a AUC pode ser utilizado para verificar se um determinado classificador é ou não adequado para o modelo utilizando os diferentes métodos já apresentados. Porém essa decisão de eliminar notas seria um fator a ser discutido posteriormente, na presente aplicação todas as notas serão consideradas independente da sua contribuição.

4.5 Modelo de Regressão Logística

A Tabela 6 apresenta as estimativas, erros padrão, estatística z e valor p associado a cada um dos coeficiente do modelo

Tabela 6 – Estimativas, erro padrão e valor-p para os parâmetros do modelo de regressão logística.

	Estimativa	Erro padrão	Estatística z	Pr($>\|z\|$)
Intercepto	0,42	0,05	8,00	<0,001
Red	0,23	0,05	4,34	<0,001
Mat	0,23	0,07	3,18	<0,001
CN	0,21	0,08	2,66	0,01
Port	0,11	0,06	1,87	0,06
CH	-0,10	0,06	-1,55	0,12

Considerando um nível de significância de 5% as variáveis Por e CH são não significativas para o modelo. Além disso para testar a existência de multicolinearidade, o fator de inflação da variância foi calculado para cada uma das variáveis, sendo apresentado na Tabela 7

Tabela 7 – Fator de Inflação da Variância

	Red	Mat	Cn	Port	CH
VIF	1,0163	1,7302	2,0541	1,1999	1,4562

O fator de inflação da variância é uma medida importante para identificar a presença de multicolinearidade no modelo que resulta em maiores erros padrão associados as estimativas dos parâmetros, sendo agravado a quanto maior forem as correlações entre as variáveis explicativas do modelo. Como informado em Kassambara (2018), valores excedendo 5 ou 10 são indicativos de presença de multicolinearidade, de acordo com esse critério, os valores presentes na Tabela 7 não informam a presença desta condição.

A avaliação da suposição de linearidade na escala do preditor linear pode ser verificada pelos gráficos na Figura 13, percebe-se que para todas as notas com exceção de Ciências Humanas a suposição parece estar sendo satisfeita.

Um modelo apenas com as variáveis significativas foi ajustado, em que o valor dos coeficientes, erros padrão, estatística z e valor p são apresentados na Tabela 8.

Figura 13 – Gráficos com relação a cada uma das variáveis explicativas na escala do preditor linear para teste de linearidade.

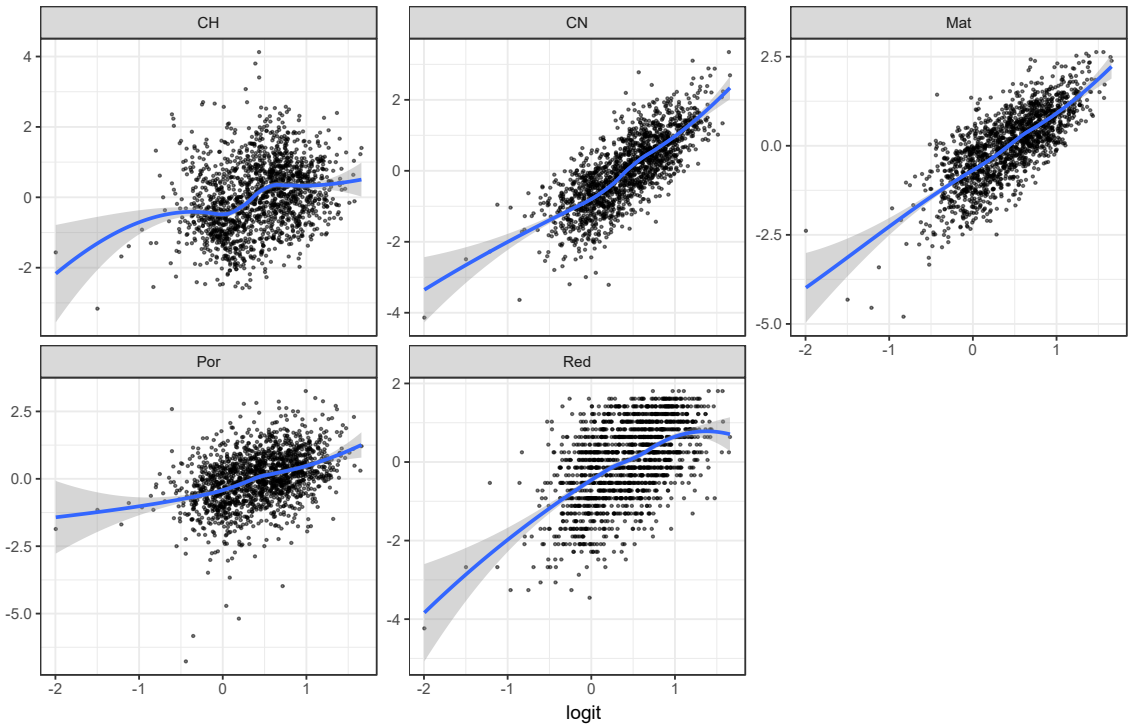


Tabela 8 – Estimativas, erro padrão e valor-p para os parâmetros do modelo de regressão logística sem as variáveis Por e CH.

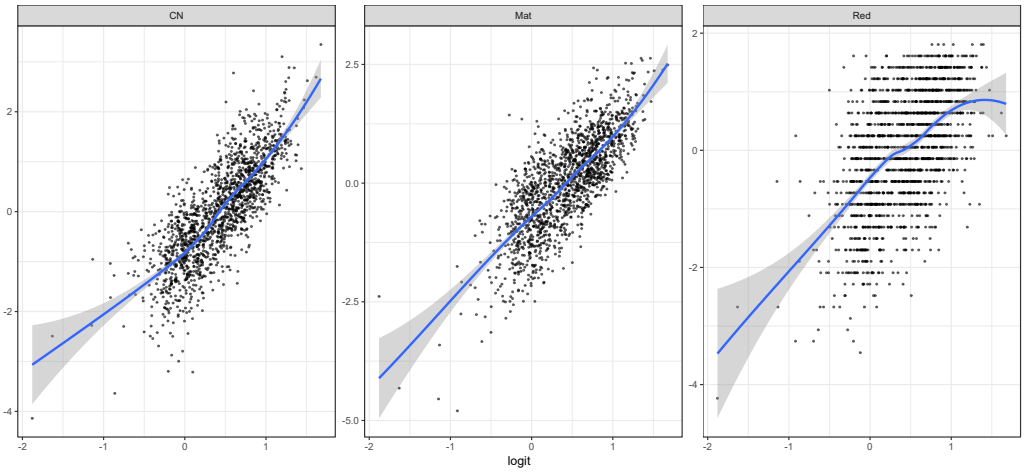
	Estimativa	Erro padrão	Estatística z	Pr($>|z|$)
(Intercepto)	0,42	0,05	7,99	0.00
Red	0,23	0,05	4,33	0,00
Mat	0,22	0,07	3,14	0,00
CN	0,19	0,07	2,71	0,01

A um nível de 5% de significância, todas as variáveis foram significativas para o modelo. O fator de inflação da variância é apresentado na Tabela 9, é perceptível que não há indícios de multicolinearidade. A Figura 14 apresenta os gráficos para teste da suposição de linearidade na escala do preditor linear, em que não a indícios de violação dessa exigência.

Tabela 9 – Fator de Inflação da Variância

	Red	Mat	CN
VIF	1,0153	1,7290	1,7337

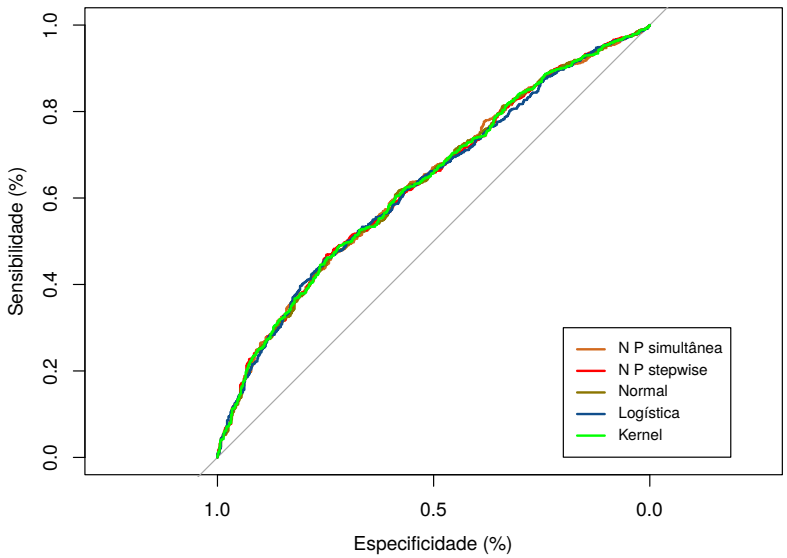
Figura 14 – Gráficos com relação a cada uma das variáveis explicativas na escala do preditor linear para teste de linearidade.



4.6 Resultados

O gráfico contento as curvas ROC de cada modelo está presente na Figura 15. É possível perceber que as curvas sugerem que todos o métodos apresentaram eficiências semelhantes.

Figura 15 – Comparativo das curvas ROC para cada um dos modelos utilizados.



Para a definição de um ponto de corte foi utilizado a função *coords* presente no pacote *pROC* (ROBIN *et al.*, 2011), nela existem alguns métodos utilizados para a escolha de um ponto de corte adequado, optou-se por utilizar aquele que resultasse na maior soma de sensibilidade e especificidade. Resultados mais precisos de cada abordagem podem ser verificados nas tabelas a seguir. Em geral todos os modelos apresentaram resultados insuficientes,

existe certo desempenho na classificação de indivíduos desistentes mas para o outro grupo de classificação a taxa de acerto é de aproximadamente 50%.

Tabela 10 – Tabela de contingência para modelo normal discriminante linear.

		Predito	
		0	1
Resposta	0	463	165
	1	499	442
ponto de corte			0,75
AUC			0,6326

Nota: sensibilidade e especificidade iguais a 0,47 e 0,74.

Tabela 11 – Tabela de contingência para modelo linear não paramétrico (*stepwise*).

		Predito	
		0	1
Resposta	0	470	158
	1	500	441
ponto de corte			0.87
AUC			0,6329

Nota: sensibilidade e especificidade iguais a 0,47 e 0,75.

Tabela 12 – Tabela de contingência para modelo linear não paramétrico (simultâneo).

		Predito	
		0	1
Resposta	0	462	166
	1	501	440
ponto de corte			0,69
AUC			0,6340

Nota: sensibilidade e especificidade iguais a 0,47 e 0,74.

Tabela 13 – Tabela de contingência para modelo de regressão logística.

		Predito	
		0	1
Resposta	0	470	158
	1	507	434
ponto de corte			0,65
AUC			0,6304

Nota: sensibilidade e especificidade iguais a 0,46 e 0,75.

Tabela 14 – Tabela de contingência para modelo linear utilizando kernel.

		Predito	
		0	1
Resposta	0	470	158
	1	522	419
ponto de corte			0,77
AUC			0,6337

Nota: sensibilidade e especificidade iguais a 0,46 e 0,75.

Portanto observa-se que para todos os modelos, a eficiência dos mesmos quando considera-se a AUC é questionável. Algo a se avaliar é com respeito a eficiência dessas metodologias comparada a de medidas mais triviais, como por exemplo a média aritmética das notas. Com base nisso, um modelo simples na forma de $W = \text{Mat} + \text{CN} + \text{Red} + \text{Port} + \text{CH}$ foi calculado. As informações sobre o resultados da aplicação do classificador simples estão presentes na Tabela 15.

Tabela 15 – Tabela de contingência para modelo linear simples.

		Predito	
		0	1
Resposta	0	452	176
	1	483	458
ponto de corte			1,13
AUC			0,6160

Nota: sensibilidade e especificidade iguais a 0,48 e 0,72.

Tomando como referência a medida AUC, é possível perceber que o modelo simples apresentou o menor valor para esta medida. Para auxiliar nas interpretações dos resultados, utilizou-se um teste de hipóteses unilateral para avaliar se a AUC obtida pelos métodos apresentados é significativamente maior do que aquela obtida pelo modelo simples. Sendo assim, seja AUC_A a medida associada ao modelo simples e AUC_B uma quantidade genérica referente a AUC de qualquer modelo que deseje comparar, o teste possui as seguintes hipóteses

$$\begin{cases} H_0: & \text{AUC}_A = \text{AUC}_B, \\ H_1: & \text{AUC}_A < \text{AUC}_B. \end{cases}$$

O teste de DeLong unilateral foi utilizado para o comparativo das AUC's, os valores da estatística z e seu valor-p para cada um dos testes está presente na Tabela 16. Percebe-se que AUC de todos os modelos apresentados é significativamente maior do que a do modelo

simples, considerando um nível de significância de 5%. Apesar disso, a melhora no desempenho de classificação continua insuficiente.

Tabela 16 – Resultados para teste de DeLong.

Comparação	Estatística z	valor-p
Normal	-2,22	0,013*
N P *stepwise*	-2,58	0,005*
N P simultânea	-2,42	0.007*
Log	-1,91	0,028*
Kernel	-2,52	0,006*

Nota: O simbolo "*"representa as relações estatisticamente significativas à 5% de significância.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foram apresentados termos associados a métodos de classificação como as medidas de eficiência, ponto de corte e curva ROC, abordando seu contexto histórico e teórico, propriedades matemáticas, diferentes processos de estimação e a variabilidade associada a mesma.

Foram discutidos também métodos destinados à combinação linear de classificadores tendo como critério de maximização da AUC ou da pAUC, com exceção do método de combinação discriminante linear para matrizes de variância e covariância proporcionais em que o resultado é a curva ROC dominante.

Todos esses métodos foram estudados tendo em mente sua aplicação direta no estudo precoce da evasão discente na Universidade Federal do Ceará, mais especificadamente no Centro de Tecnologia. Em que foram consideradas as notas do Exame Nacional do Ensino Médios dos ingressantes via processo seletivo ENEM/SiSU dos de 2014 e 2015.

Na análise da possibilidade de aplicar pesos nas notas do Exame, assim como na definição das notas mínimas para o possível ingresso, que poderia ser adaptada ao sistema de seleção da Universidade (ENEM/SiSU), a fim de aprimorar o processo seletivo, identificando previamente os indivíduos com menor propensão à evasão, foi visto que a solução para esse problema pode ir muito além das notas no ENEM.

No trabalho é percebido que a eficiência para todos os modelos de classificação foi insatisfatória, sugerindo que para um critério de combinação linear, as notas do ENEM não são bons classificadores, mesmo considerados conjuntamente sob o critério da AUC, em relação à evasão discente nos cursos do Centro de Tecnologia.

Outras metodologias mais sofisticadas podem ser aprimoradas para uma obtenção de resultados melhores, mas isso foge da possibilidade da implementação no sistema de seleção via ENEM/SiSU, assim como da proposta desta monografia.

Para trabalhos futuros outros fatores poderiam ser considerados tais como a utilização da pAUC para um estudo dos custos que cada erro de classificação pode resultar servindo assim para estabelecer um intervalo viável para a especificidade. Além disso, seria interessante uma análise mais aprofundada considerando a política de cotas ou até mesmo a utilização da variável renda que talvez seja determinante para a permanência no curso para uma parcela considerável dos alunos. Em sendo estes fatores realmente significativos o problema passaria a ser mais do que o desenvolvimento de uma métrica de seleção e mais a identificação de alunos sujeitos a

evasão que poderiam ser auxiliados por medidas assistenciais adotadas pelas instituições, nesse caso, modelos mais sofisticados poderiam ser utilizados como os de aprendizagem de máquina.

REFERÊNCIAS

ANDERSON, T. W.; BAHADUR, R. R. *et al.* Classification into two multivariate normal distributions with different covariance matrices. **The annals of mathematical statistics**, Institute of Mathematical Statistics, v. 33, n. 2, p. 420–431, 1962.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, Oxford University Press, v. 68, n. 1, p. 13–20, 1981.

AZZALINI, A.; BOWMAN, A. W. A look at some data on the old faithful geyser. **Applied Statistics**, JSTOR, p. 357–365, 1990.

BAMBER, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. **Journal of mathematical psychology**, Elsevier, v. 12, n. 4, p. 387–415, 1975.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 211–252, 1964.

BRAGA, A. **Curvas ROC: aspectos funcionais e aplicações**. Tese (Doutorado) — Universidade do Minho, BRAGA, 2001.

COCHRAN, W. G. **Sampling Techniques**. 3. ed. New York: John Wiley & Sons, 1977.

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, JSTOR, p. 837–845, 1988.

DODD, L. E.; PEPE, M. S. Partial auc estimation and regression. **Biometrics**, Wiley Online Library, v. 59, n. 3, p. 614–623, 2003.

EGAN, J. P. **Signal detection theory and ROC-analysis**. New York: Academic Press, 1975.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

GREEN, D. M.; SWETS, J. A. **Signal detection theory and psychophysics**. New York: Wiley, 1966.

HANLEY, J. A.; HAJIAN-TILAKI, K. O. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. **Academic radiology**, Elsevier, v. 4, n. 1, p. 49–58, 1997.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HANLEY, J. A.; MCNEIL, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. **Radiology**, v. 148, n. 3, p. 839–843, 1983.

HSU, M.-J.; HSUEH, H.-M. The linear combinations of biomarkers which maximize the partial area under the roc curves. **Computational Statistics**, Springer, v. 28, n. 2, p. 647–666, 2013.

JIANG, Y.; METZ, C. E.; NISHIKAWA, R. M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. **Radiology**, v. 201, n. 3, p. 745–750, 1996.

KASSAMBARA, A. **Machine Learning Essentials: Practical Guide in R**. 1. ed. [S.l.]: CreateSpace Independent Publishing Platform, 2018.

LEHMANN, E. L. Consistency and unbiasedness of certain nonparametric tests. **The Annals of Mathematical Statistics**, JSTOR, p. 165–179, 1951.

LIU, A.; SCHISTERMAN, E. F.; ZHU, Y. On linear combinations of biomarkers to improve diagnostic accuracy. **Statistics in medicine**, Wiley Online Library, v. 24, n. 1, p. 37–47, 2005.

LLOYD, C. J. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 93, n. 444, p. 1356–1364, 1998.

LOBO, R. A evasão no ensino superior brasileiro – novos dados. **Estadão**, São Paulo, 7 out. 2017. Disponível em: <https://educacao.estadao.com.br/blogs/roberto-lobo/497-2/>. Acesso em: 08 jan. 2019.

LOBO, R.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

LUSTED, L. B. Logical analysis in roentgen diagnosis: memorial fund lecture. **Radiology**, The Radiological Society of North America, v. 74, n. 2, p. 178–193, 1960.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, JSTOR, p. 50–60, 1947.

MCCLISH, D. K. Analyzing a portion of the roc curve. **Medical Decision Making**, Sage Publications Sage CA: Thousand Oaks, CA, v. 9, n. 3, p. 190–195, 1989.

MEC/SESU. **Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras**. Brasília: ANDIFES/ABRUEM/SESU/MEC. 1996.

MOOD, A. M. **Introduction to the Theory of Statistics.** [S.l.]: McGraw-hill, 1950.

OLIVARES, Viviane da Silva. **A curva ROC e suas aplicações**. 2009. Monografia (Graduação em Matemática Aplicada e Computacional) − Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2009.

PEPE, M. S. **The statistical evaluation of medical tests for classification and prediction**. [S.l.]: Oxford University Press, 2003. v. 28.

PEPE, M. S.; THOMPSON, M. L. Combining diagnostic test results to increase accuracy. **Biostatistics**, Oxford University Press, v. 1, n. 2, p. 123–140, 2000.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>.

ROBIN, X.; TURCK, N.; HAINARD, A.; TIBERTI, N.; LISACEK, F.; SANCHEZ, J.-C.; MÜLLER, M. proc: an open-source package for r and s+ to analyze and compare roc curves. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 77, 2011.

SEN, P. K. On some convergence properties of ustatistics. **Calcutta Statistical Association Bulletin**, SAGE Publications Sage India: New Delhi, India, v. 10, n. 1-2, p. 1–18, 1960.

SU, J. Q.; LIU, J. S. Linear combinations of multiple diagnostic markers. **Journal of the American Statistical Association**, Taylor & Francis, v. 88, n. 424, p. 1350–1355, 1993.

SWETS, J.; PICKETT, R. Evaluation of diagnostic systems: Methods from signal detection theory academic. **New York**, 1982.

THOMPSON, M. L.; ZUCCHINI, W. On the statistical analysis of roc curves. **Statistics in Medicine**, Wiley Online Library, v. 8, n. 10, p. 1277–1290, 1989.

WAND, M. P.; JONES, M. C. **Kernel smoothing**. [S.l.]: Chapman and Hall/CRC, 1994.

WELCH, B. L. Note on discriminant functions. **Biometrika**, JSTOR, v. 31, n. 1/2, p. 218–220, 1939.

YAN, Q.; BANTIS, L. E.; STANFORD, J. L.; FENG, Z. Combining multiple biomarkers linearly to maximize the partial area under the roc curve. **Statistics in medicine**, Wiley Online Library, v. 37, n. 4, p. 627–642, 2018.

ZOU, K. H.; TEMPANY, C. M.; FIELDING, J. R.; SILVERMAN, S. G. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral ct of ureteral stones. **Academic radiology**, Elsevier, v. 5, n. 10, p. 680–687, 1998.

APÊNDICE – FIGURA

Figura 16 – Matriz de gráficos para variáveis referentes as provas do ENEM. Cor: Coeficiente de correlação de Pearson, D: Alunos Desistentes, N.D: Alunos Não Desistentes.