1 Introdução

A indústria de jogos eletrônicos tem evoluído rapidamente, uma atividade que até pouco tempo era vista como diversão e passatempo de crianças e jovens, torna-se algo muito mais sério. O número de jogadores profissionais neste meio tem aumentado de modo visto. Tais fatos são comprovados pela crescente divulgação de torneios mundiais e grande quantidade de times profissionais participantes. Um termo que se ouve com maior frequência é "*eSports*"(Esporte Eletrônico), já o próprio nome faz referência aos jogos eletrônicos, porém apresentando-os como verdadeiro esporte que possui: seus atletas profissionais, disputas acirradas em nível mundial, equipes favoritas e uma crescente audiência. Nota-se também o aumento significativo dos valores das premiações dos campeonatos promovidos neste meio, encontrando-se na casa de milhões de dólares.

Dentre a gama de jogos eletrônicos, o presente trabalho abordará o Dota2, um jogo em que duas equipes se enfrentam num campo de batalha. As equipes são formardas por 5 jogadores e cada jogador escolhe um personagem (Herói) para controlar dentre 111 disponíveis, ou seja, apenas 10 personagens participam da partida, não havendo escolhas repetidas. Cada um dos personagens possui características distintas em relação aos demais (habilidades, poderes e atributos), tornando o jogo muito diversificado. O objetivo do jogo consiste em destruir a estrutura principal da base da equipe inimiga. A equipe que realizar primeiro tal feito é considerada vencedora, não existindo tempo limite de jogo.

Atualmente o número de jogadores únicos registrados no jogo Dota2 ultrapassa 12 milhões, sendo que os picos diários de acesso são supriores a 900 mil jogadores conectados simultaneamente (Dota2, 2016). Dota2 chamou a atenção do mundo quando a empresa responsável por seu desenvolvimento, *Valve*, anunciou o *The International* em 2011, torneio mundial possuindo o valor da premiação recorde até então, 1 milhão de dólares para equipe vencedora. Em 2015, a premiação total do *The International* foi superior a 18 milhões de dólares, sendo U$6.634.661,00 do montante total, designado para a equipe vencedora. Até a presente data não há nenhum torneio neste segmento que o supere em premiação. Ainda que Dota2 não seja o *eSport* que possui maior número de campeonatos, é o jogo eletônico que possui o maior valor de premiações acumuladas, 59 milhões de dólares em 587 torneiros (EsportsEarnings, 2016). É mais que notório a dimensão de como *eSports* se tornou algo muito sério e profissional, sendo necessário dedicação de tempo e grande investimento financeiro.

Logo, equipes ao redor do mundo se preparam durante o ano, buscando sempre melhorar seu desempenho e estudando o jogo minuciosamente. A escolha dos personagens pode favorecer determinada equipe em relação à outra, ou seja, é importante levar em consideração quais personagens tem boa interação uns com os outros, como também, suas vantagens em relação aos personagens selecionados pela equipe adversária. O placar da partida apresenta quantas mortes cada equipe realizou sobre a equipe inimiga, o que é um forte indicativo de quem está mais próximo da vitória. Entretanto, não é a única medida a ser avaliada, havendo também o desenvolvimento individual dos personagens, como: "ouro por minuto", "experiência por minuto", "número de mortes realizadas", "número de mortes sofridas", dentre outras. Nota-se então, uma grande quantidade de dados que podem ser coletados, processados e utilizados de cada partida para uma futura ou imediata tomada de decisão. Modelos estatísticos e algoritmos computacionais devem ser empregados para obtenção de informações precisas que auxiliarão desde a etapa de seleção dos personagens, como também desempenhos almejados durante a partida, sempre em busca de maximizar a probabilidade de vitória da equipe.

A utilização de métodos de Aprendizado de Máquina (AM) tem permitido resolver problemas cada vez mais complexos. Embora sua fundamentação teórica não seja tão recente, o avanço da tecnologia permitiu tanto sua aplicação quanto escalabilidade (Faceki et al., 2015). O presente trabalho abordará duas técnicas de AM: *Naive Bayes* (NB) e *Tree Augmented Naive Bayes* (TAN) (Friedman et al., 1997). Ambas contempladas pela teoria de Redes Bayesianas para classificação, que a partir de variáveis auxiliares atualiza a probabilidade de determinada classe (categoria, nível, evento) da variável resposta por meio do Teorema de Bayes. Tais técnicas possuem o objetivo similar à Análise Discriminante ou Regressão Logística do ponto de vista estatístico, pretendendo-se classificar determinado objeto ou indivíduo a partir de suas características.

2 Objetivo

O objetivo principal do presente trabalho consiste em utilizar as técnicas de AM citadas anteriormente para classificar determinada equipe como vencedora ou perdedora, segundo variáveis que podem ser observadas do início ao final da partida.

Como objetivo secundário, utilizando os modelos desenvolvidos, será criada uma inter-

face para qualquer usuário inserir dados, fornecendo em tempo real respostas para alguns questionamentos levantados por entusiastas, jogadores e treinadores:

- Qual conjunto de personagens tem vantagem sobre outro conjunto?
- Quais personagens são melhores caso a partida se alongue muito?
- Qual é o rendimento ideal para um determinado personagem (ouro e experiência por minuto)?
- Qual dentre os jogadores da minha equipe deve melhorar o desempenho durante a partida?
- Em qual equipe eu deveria apostar?

3 Justificativa

Apesar da área de *eSports* ser crescente como foi apresentado anteriormente, os trabalhos encontrados que abordam o tema utilizam apenas uma variável para predição de partidas de Dota2: "Seleção de personagem". Ou seja, dentre todas as informações que podem ser utilizadas, apenas uma tem sido levado em conta, quais personagens participam da partida. (Conley, 2013)

Acredita-se que o incremento de outras variáveis aumentarão as taxas de assertividade, fornecendo maior convicção para tomada de decisão. Possibilita-se então a utilização de dados concretos e análises mais precisas para: corrigir erros, aprimorar jogadas, analisar seus jogadores, identificar melhorias a serem realizadas, oportunidades que devem ser aproveitadas, bem como uma infinidade de ações durante a partida, que consequentemente melhorarão o desempenho da equipe.

Após a conclusão do trabalho final, caso os resultados atinjam às expectativas, o mesmo poderá ser utilizado para outros jogos eletrônicos que estejam no cenário competitivo e possuam seus dados de forma aberta bem como ocorre em Dota2.

O presente trabalho busca desde já contribuir com a sociedade científica onde o autor está inserido, apresentando uma técnica que dificilmente é abordada na graduação em Estatística. Com isso, aumentar o leque de métodos e técnicas que podem ser adotados para soluções de problemas na área.

4 Metodologia

Para desenvolvimento do trabalho o mais corretamente possível, tanto pelo ponto de vista da Estatística quanto computacional, serão definidas algumas etapas a serem seguidas:

A coleta de dados é uma das etapas de maior complexidade. Será utilizada uma API (*Application Programming Interface*) fornecida pela empresa *Valve*, que possibilita a coleta de partidas públicas armazenadas em seus servidores. A coleta será feita em tempo real e de maneira sistemática, a cada 10 minutos um algoritmo busca pelas 100 últimas partidas encerradas, armazenando-as em um banco de dados local. Em coletas preliminares, cerca de 8 mil partidas foram coletadas por dia, assim, ao final do trabalho o número de observações chegará na casa dos milhões, visto que esta etapa será realizada ao longo de todo trabalho. Pela busca de partidas em nível profissional, serão coletadas apenas as partidas dos jogadores classificados como *"very high–level"* ("nível muito alto", índice definido pelo próprio jogo), o que representa 11,9% dentre todos jogadores, sendo estes os que possuem os melhores desempenhos em Dota2. Partidas abaixo deste nível podem comprometar a qualidade do dados para realização do estudo, devido sua heterogeneidade.

A construção da base de dados é etapa intrínseca à anterior. A base de dados deve ser consistente e alinhada com o propósito do estudo. Definindo assim, as variáveis que irão compô-la desde o início da coleta. As seguintes variáveis são necessárias para o estudo: equipe vencedora, personagens em jogo e a qual equipe pertence, experiencia por minuto e ouro por minuto de cada personagem, tempo de duração da partida, ID da partida.

Na etapa da modelagem, é conveniente que seja proposto um modelo teórico baseado nas técnicas propostas, facilitando o entendimento de como será realizado em vias computacionais. Nesta etapa será tomado todo cuidado necessário para respeitar notações matemáticas, bem como propriedades da teoria de probabilidade. É possível apresentar modelos na forma analítica para as duas técnicas que serão utlizadas, diferentemente de outras técnicas de AM, em que são etapas computacionais e não uma fórmula fechada.

A implementação computacional, é considerada como mais extensa e intensiva etapa. Para a técnica NB, a linguagem de programação *Python* fornece o método implementado em uma de suas bibliotecas (*scikit-learn*), que será muito útil. Porém, para a técnia TAN, não existe pacote ou *software* estatístico que a contenha implementada da meneira que se pretende aplicá-la, logo, será necessário todo seu desenvolvimento pelo próprio autor.

Faz parte do processo de modelagem a realização de validações dos modelos propostos, verificando suas performances segundo métricas, partições e amostragens da base de dados que ainda serão estudadas. Toda etapa computacional será realizada na linguagem de programação *Python* com auxilio de suas bibliotecas (*pandas*, *numpy*, *scikit-learn*).

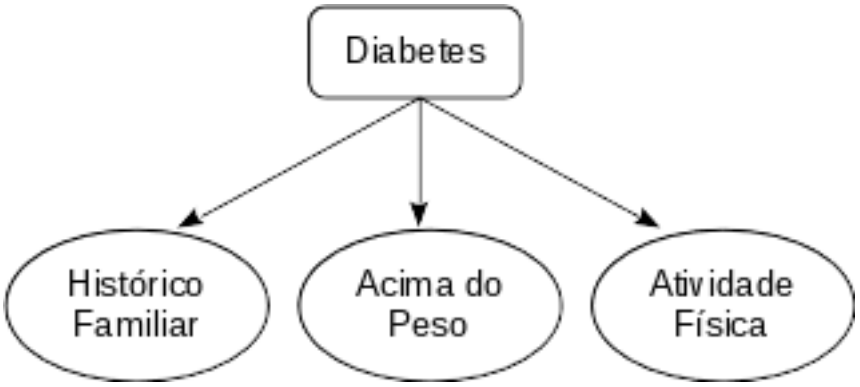
Para ilustrar como aplica a técnica NB, é apresentado seguir um exemplo fictício detalhando todos os passos para a resolução do problema. Além da formulação teórica do ponto de vista probabilístico, há também um modelo realizado utilizando representação de grafos, auxiliando no entendimento da técnica.

***Exemplo**:* Um pesquisador pretende realizar um diagnóstico de diabetes em um de seus pacientes sem consultar exames de laboratório. Para isso, ele utilizará a base de dados de seus pacientes históricos, onde há os respectivos diagnóstico e informações pessoais. A base de dados do pesquisador é apresentada pela tabela 1.

Tabela 1: Base de dados de pacientes

Paciente	Diabetes (Y)	Histórico Familiar (X_1)	Acima do Peso (X_2)	Atividade Física (X_3)
1	1	1	1	1
2	1	1	1	0
3	1	0	1	0
4	1	1	0	0
5	0	1	0	1
6	0	0	1	1
7	0	0	0	0

Figura 1: Esquema de grafos da técnica Naive Bayes



Fonte: Produzida pelo autor

Ao interpretar o problema do pesquisador, define-se a variável de interesse (resposta) como sendo "Diabetes", logo, todas outras variáveis do conjunto de dados são consideradas auxiliares (regressoras, explicativas, etc). Desta forma, a figura 1 representa a modelagem via NB por meio de grafos, em que todas variaveis auxiliares são independentes umas das outras ao serem condicionadas a Y (resposta). Isto é, todas as variáveis auxiliares são independentes entre sí dado a ocorrência de determinado nível (classe) na variável resposta, propriedade conhecida como independência condicional em probabilidade.

Ao analisar a tabela 1, é fácil verificar que a probabilidade de uma pessoa qualquer ser diagnosticada com diabetes é $4/7$, isto é, $p(Y = 1) = 4/7$. É desejável atualizar essa probabilidade condicionando a ocorrência de y à $\mathbf{x} = (x_1, x_2, x_3)^t$. Em outras palavras, as variáveis auxiliares fornecerão mais informação para aumentar a certeza do diagnóstico. O Teorema de Bayes pode ser utilizado para esta tarefa:

$$p(y|\mathbf{x}) = \frac{p(y)\prod_{i=1}^{3} p(x_i|y)}{p(\mathbf{x})} \tag{1}$$

Ao utilizar o método NB, pressupõem-se que as variáveis auxiliares são independentes dado uma classe específica de Y, pode-se então realizar o seguinte passo:

$$p(x_1, x_2, x_3|y) = p(x_1|y)p(x_2|y)p(x_3|y) = \prod_{i=1}^{3} p(x_i|y)$$

O critério de classificação criado utiliza uma razão de chances para decidir se o diagnóstico será "diabetes positiva"ou não. Quando esta razão de chances for maior que 1, classifica-se como diabético:

$$\frac{p(Y = 1|x_1, x_2, x_3)}{p(Y = 0|x_1, x_2, x_3)} = \frac{p(Y = 1)\prod_{i=1}^{3} p(x_i|Y = 1)}{p(Y = 0)\prod_{i=1}^{3} p(x_i|Y = 0)} \geqslant 1 \Rightarrow \text{Diabetes}$$

Tal artifício nos livra da necessidade do cálculo de $p(x_1, x_2, x_3)$, tendo em vista que não seria trivial definir a distribuição conjunta de tais variáveis, uma vez que as mesmas não são independentes entre sí.

Devido a forma como as variáveis foram definidas (binárias), é razoável atribuir a distribuição de probabilidade Bernoulli a cada uma delas. Porém para cada $x_i|y$ tem-se uma distribuição distinta, devido a isso, a notação fica mais carregada que o usual:

$$p(X_i = x_i|Y = y) = {\theta_{iy}}^{x_i}(1 - \theta_{iy})^{1-x_i}, \text{ para } x_i \in \{0, 1\} \; ; \; 0 < \theta_{iy} < 1$$

Onde θ_{iy} é o parâmetro da distribuição da i-ésima variável, dado uma classe específica de Y. Ou seja, a distribuição de cada variável auxiliar se altera dado a ocorrência de Y.

Estima-se θ_{iy} por meio de $\hat{\theta}_{iy}$, definido da seguinte forma:

$$\hat{\theta}_{iy} = \frac{\sum_{j=1}^{n_y} x_{ij}}{n_y}$$

Em que n_y é o tamanho da amostra na classe y, isto é, quando y ocorre. É possível agora estimar todos os parâmetros necessários e apresentar o modelo ajustado. A seguir são apresentadas as estimativas dos parâmetros que irão compor o modelo.

$$\hat{\theta}_1 = \begin{cases} 1/3 & \text{para } y = 0; \\ 3/4 & \text{para } y = 1. \end{cases} ; \quad \hat{\theta}_2 = \begin{cases} 1/3 & \text{para } y = 0; \\ 3/4 & \text{para } y = 1. \end{cases} ; \quad \hat{\theta}_3 = \begin{cases} 2/3 & \text{para } y = 0; \\ 1/4 & \text{para } y = 1. \end{cases}$$

É definido o seguinte modelo para o diagnóstico de diabetes:

$$\frac{p(Y=1|x_1,x_2,x_3)}{p(Y=0|x_1,x_2,x_3)} = \frac{(4/7)(3/4)^{x_1}(1/4)^{1-x_1}(3/4)^{x_2}(1/4)^{1-x_2}(1/4)^{x_3}(3/4)^{1-x_3}}{(3/7)(1/3)^{x_1}(2/3)^{1-x_1}(1/3)^{x_2}(2/3)^{1-x_2}(2/3)^{x_3}(1/3)^{1-x_3}} \quad (2)$$

Vale lembrar que ao realizar predições, caso (2) resulte um valor maior que 1, classifica-se o paciente como diabético. Para um caso em que o paciente possui histórico familiar, não está acima do peso e não pratica atividade física, tem-se $\mathbf{x} = (1, 0, 0)^t$ e substituindo os valores em (2):

$$\frac{p(Y=1|\mathbf{x})}{p(Y=0|\mathbf{x})} = \frac{0.08}{0.03} = 2.67 \geqslant 1 \Rightarrow \text{Diabetes}$$

Portanto, o paciente é diagnosticado como diabético. O pesquisador pode recorrer a (2) sempre que não conseguir realizar o exame de laboratório, utilizando apenas as variáveis fornecidas pelo paciente. ■

Para aplicação da técnica NB na perspectiva de Dota2, as variáveis utilizadas para modelagem são referentes a presença ou não dos personagens na partida. Desta forma, cada x_i representa a presença ou não do i-ésimo personagem.

Embora não seja apresentado um exemplo de como a técnica TAN é aplicada, a mesma é considerada como uma expansão de NB. Uma vez que NB supõem independência entre todas as variáveis dado y, já para TAN, pode-se supor que uma (ou mais) variável é dependente de uma e somente uma outra variável (Faceki et al., 2015).

5 Cronograma

Etapa	Abr.	Maio	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
1.	X	X	X	X	X				
2.	X								
3.	X	X	X	X	X	X	X	X	
4.	X	X	X	X					
5.					X				
6.			X	X	X	X	X		
7.								X	X

1. Levantamento Bibliográfico

2. Estruturação da Base de Dados

3. Coleta de Dados

4. Modelagem Naive Bayes

5. Entrega Relatório Parcial

6. Modelagem Tree Augmented Naive Bayes

7. Entrega Relatório Final

Bibliografia

Bolfarine, H. (2010). *Introdução à inferência estatística*. SBM, Rio de Janeiro, 2a edição.

Conley, K. (2013). How Does He Saw Me? A Recommendation Engine for Picking Heroes in Dota 2. Technical report, Stanford University.

Dota2 (2015). The International 2015. Disponível em: `<http://www.dota2.com/international/overview>`. Último Acesso: 12 de abril de 2016.

Dota2 (2016). Disponível em: `<http://www.dota2.com>`. Último Acesso: 12 de abril de 2016.

Downey, A. B. (2012). *Think Bayes: Bayesian Statistics Made Simple*. Green Tea Press.

EsportsEarnings (2016). Disponível em: `<http://www.esportsearnings.com>`. Último Acesso: 12 de abril de 2016.

Faceki, K. et al. (2015). *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. LTC.

Friedman, N. et al. (1997). Bayesian network classifiers. *Machine Learning*, 29:131–163.

McKinney, W. (2013). *Python for Data Analysis*. O'Reilly Media Inc.

Sheldon, R. (2010). *Probabilidade: Um curso moderno com aplicações*. Bookman, Porto Alegre, 8a edição. Tradução: Alberto Resende De Conti.

Summerfield, M. (2012). *Programação em Python 3: uma introdução completa à linguagem Python*. Alta Books, Rio de Janeiro.

Zhang, H. (2004). The Optimality of Naive Bayes. *University of New Brunswick*.