

A Previsibilidade em Eventos Aleatórios

- Estudo de Caso -

Aluna: [Silvia Puetter Mattos](#)

Orientadora: [Manoela Kohler](#)

Trabalho apresentado ao curso BI MASTER como pré-requisito para conclusão de curso e obtenção de crédito na disciplina "Projetos de Sistemas Inteligentes de Apoio à Decisão".

- [Códigos em arquivo anexo neste projeto.](#)
-

Dedicatória

A todos que de alguma forma seguem alimentando e impulsionando a minha paixão pelas ciências e pelo conhecimento, dedico este estudo. Àqueles que seguem questionando na busca incansável dos 'porquês'. Professores queridos de toda minha trajetória acadêmica. Aldo, Inge, Giselle, Joaquim, Leo.

Agradecimentos

À querida amiga, Vitoria Vellozo, agradeço o compartilhamento de tanto conhecimento sempre e pelo incentivo persistente e fundamental ao meu crescimento intelectual e pessoal.

À professora e orientadora, Manoela Kohler, pela didática impecável, irretocável, indescritível. Pela dedicação, ajuda e paciência na elaboração do estudo.

À toda equipe docente do curso BI MASTER – PUC Rio, por sua maestria na disseminação de ciência e aprendizado.

Resumo

O sucesso de uma tomada de decisão diante de incerteza é uma habilidade rara e muitas vezes aperfeiçoada através da experiência do indivíduo. A intuição humana é usada quando se desconhece fatos e dados. No entanto, quando os dados existem, a evidência existe e não há motivos para ignorá-la. Um método utilizado para quantificar a incerteza e ajudar a tomada de decisão é com o uso das probabilidades e toda matemática e raciocínio envolvido por trás dos panos. A partir da revolução tecnológica, e com o grande volume de dados que o big data aflorou, a linha tênue entre evidência e intuição tem se estreitado cada vez mais, abrindo espaço para novas abordagens. Desde quando os algoritmos de aprendizado transformaram as torrentes de dados em novo conhecimento científico, novas explorações foram possibilitadas. O trabalho se propõe a, através de um levantamento na literatura acerca da aleatoriedade, associado à construção de um modelo de probabilidade – modelo preditivo de machine learning - analisar os fenômenos aleatórios do processo de sorteio das dezenas do jogo lotérico Mega Sena.

Abstract

Successful decision making in the face of uncertainty is a rare skill that is often honed through individual experience. Human intuition is often used when facts and data are still unknown. However, when the data is there, the evidence is also there and there is no reason to ignore it. One method used to quantify uncertainty and help decision making is use probabilities and all the math and knowledge involved behind the scenes. Since the technological revolution, and with the large volume of data that big data has emerged, the fine line between evidence and intuition has increasingly narrowed, opening space for new findings. Ever since learning algorithms transform torrents of data into new scientific knowledge, new explorations have been made possible. This study proposes, through a survey in the literature about randomness, associated with the construction of a probability model - predictive model of machine learning - to analyze the random phenomena of the process of drawing dozens of the Mega Sena lottery game.

1. Introdução

Eventos aleatórios são eventos que ocorrem ao acaso. Embora os fenômenos aleatórios, no longo prazo, tendam a se estabilizar em padrões consistentes e previsíveis, é impossível prever resultados individuais, não importando quantas vezes o mesmo experimento tenha se repetido.¹ O grau de incerteza associado é tão alto que torna a sua possibilidade de previsão de intensa dificuldade.

Os modelos preditivos são um ramo da ciência da computação que abarca o uso de dados e algoritmos para geração de informação, e sua essência é a previsão. A predição usa as informações que se tem, denominadas de dados, para gerar aquelas que não existem (ainda). Os algoritmos de aprendizado transformam, portanto, as torrentes de dados existentes em novo conhecimento científico.^{2,3} De acordo com definição publicada no site da SAS, empresa pioneira em Business Intelligence:

*“O aprendizado de máquina (em inglês, machine learning) é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.”*⁴

Neste contexto, afloram questionamentos acerca de assunto tão longínquo na história do homem. Um mix de 3 elementos em uma única suspeição: (1) O desejo de se decifrar o futuro e obter previsões, (2) associado à matemática avançada e todos os seus instrumentos incluindo máquinas preditivas e algoritmos robustos, (3) e os dados. Muitos dados. Com o advento do big data e o tsunami de informações e registros disponíveis, a potência dos algoritmos do machine learning e todo arcabouço computacional disponível, a imprevisibilidade dos eventos aleatórios poderia ser colocada em xeque?

A exploração dessa onda de informações, atrelada ao desenvolvimento de modelos possíveis buscou procurar respostas. Ou ainda: outras perguntas e novas hipóteses. E, certamente, questionar. Sabe-se que o estudo do comportamento de um próprio fenômeno observado pode ser usado como base do estudo para esses modelos. E, o jogo da Mega Sena, além de bastante popular, é o maior jogo lotérico do Brasil. Seu histórico de resultados é representativo e seu racional de realização reúne eventos aleatórios e muitas possíveis análises e relações. Colher e explorar dados, buscar eventuais padrões, percorrer modelos e buscar considerações e questionamentos para o assunto foi a motivação do presente trabalho.

O estudo não se propõe a prever as dezenas sorteadas. Seu objetivo é o de: (a) aprofundar o conhecimento acerca da aleatoriedade e seus fatores determinantes; (b) explorar o ambiente do big data, estimulando o desenvolvimento de práticas de localização e uso da informação (extração de dados, data mining, e análise exploratória); (c) desenvolver e testar modelos de machine learning com capacidades preditivas, bem como a qualidade dos dados disponíveis; e (d) provocar reflexão quanto ao alcance dos algoritmos e o poder do big data na previsibilidade de eventos em cenários de incerteza no mundo atual.

2. Modelagem

O jogo da Mega Sena é a maior modalidade lotérica no Brasil, tendo sido criado em 1996 pela Caixa Econômica Federal. O jogo consiste em, dentro de 60 números, o apostador acertar os 6 números sorteados. Ainda são contemplados no prêmio aqueles que acertarem 5 ou 4 dezenas sorteadas, acertando a chamada quina e quadra, respectivamente. As apostas podem ser realizadas escolhendo-se de 6 a 15 dezenas, e o valor dos jogos cresce conforme se elevam as chances de acerto. Os sorteios iniciaram sendo realizados uma vez por semana, sendo atualmente realizados às quartas e aos sábados, sempre às 20 horas.⁵

De acordo com dados históricos, até 2009 os sorteios eram realizados em globos duplos e os números eram sorteados em dígitos separados, que formavam uma dezena (um número decimal de dois dígitos), de 01 a 60. Desde o sorteio n.º 1 140, de 31 de dezembro de 2009, os dois globos, originariamente de metal, foram substituídos por um único globo de acrílico, com bolas numeradas de 01 a 60, todas coloridas, uma cor para cada dezena. À título do estudo, esta alteração no processo do sorteio foi desprezada.⁴

O sorteio das dezenas segue a dinâmica de eventos aleatórios, e a probabilidade de um jogo ser sorteado é a mesma de qualquer outra combinação. A possibilidade de uma eventual previsibilidade nesse sorteio foi o centro do problema em estudo.

O racional de construção do modelo não se limitou ao histórico dos resultados do sorteio. Buscou-se ampliar as bases históricas para registros que pudessem ter relações, ainda que especuladas neste estudo, com o resultado. A ideia, portanto, foi de criar um algoritmo que estudasse o comportamento das variáveis levantadas e suas relações com os eventos finais, indagando a existência, ou não, de alguma previsibilidade.

2.1. Extração de dados

2.1.1. Histórico dos sorteios ⁵

A etapa inicial do trabalho consistiu no levantamento do histórico de todos os registros de sorteios da Mega Sena realizados desde sua criação. A Caixa Econômica Federal disponibiliza essa informação em seu site, havendo dados acerca da data do sorteio, das 6 dezenas sorteadas em sua ordem, bem como: quantidade de ganhadores da mega sena, da quina e da quadra; os valores do rateio do prêmio, a cidade de realização do sorteio, valor arrecadado, estimativas e valores para o próximo sorteio, se acumulou ou não, se foi um sorteio comum ou especial (ex: Mega da Virada) e observações quando pertinentes. Foram extraídos todos os resultados até a data de 24/05/2022.

Para o modelo a ser desenvolvido foram selecionadas apenas as informações: data do sorteio e dezenas. As demais informações foram descartadas. Embora a cidade de realização pudesse contribuir para um modelo futuro, a escassez de registros deste atributo na base inviabilizou o seu uso.

2.1.2. Histórico do clima ⁶

Tendo-se o histórico dos sorteios, levantou-se possíveis variáveis que pudessem ter alguma influência no processo do sorteio. Aqui a abordagem das possibilidades seguiu uma lógica meramente especulativa, uma vez que haja infinitas e diversas condições e propriedades que determinem e caracterizem o momento em que o sorteio é realizado.

Explorou-se a disponibilidade de dados públicos na web, de modo que um levantamento destes atributos fosse suficiente para suportar os modelos preditivos a serem propostos. Buscou-se contemplar tanto aspectos objetivos, como vertentes mais subjetivas conforme descrição a seguir.

O Instituto Nacional de Meteorologia – INMET, do Ministério da Agricultura, Pecuária e Abastecimento, dispõe de um banco de dados meteorológicos cujo histórico cobre o período dos registros históricos da Mega Sena.⁶ Foram selecionados, então, dados diários, das estações convencionais, com abrangência nacional, no período de 01/01/1996 até 24/05/2022. Foram selecionadas todas as capitais dos 26 estados e Distrito Federal. As variáveis apuradas foram: (1) evaporação do piche diária(mm), (2) insolação total diário, (3) precipitação total diário(mm), (4) temperatura máxima diária (°c), (5) temperatura média compensada diária (°c), (6) temperatura mínima diária (°c), (7) umidade relativa do ar média diária (%), (8) vento velocidade média diária (m/s).

Em cada capital, foi selecionada, dentre as estações operantes, aquela com o maior número de registros no período. Nos casos em que as estações da capital apresentavam histórico bastante limitado, com estações inoperantes por períodos superiores a 2 anos, foi selecionada outra estação no estado, que tivesse uma cobertura maior, ainda que fora da capital. Houve, ainda, situação em que a ausência de dados era bastante constante no histórico disponibilizado pelo INMET, o que fez com que alguns dados fossem mesclados com diferentes estações a fim de se alcançar uma base melhor para alimentação do modelo, posteriormente. A lista de estações contempladas no estudo pode ser conhecida nos arquivos do trabalho.

Mesmo com todas essas premissas para extração da maior base histórica possível, observou-se que alguns parâmetros a serem utilizados no modelo continuavam apresentando elevada frequência de valores nulos e, embora pudessem ser relevantes para o modelo, qualquer tratamento realizado previamente poderia não ter ganho representativo. Foi definido que variáveis em que a frequência de valor nulo fosse maior ou igual a 600, dentro de um universo de 2.472 registros possíveis, seriam descartadas da base. A métrica foi manter apenas as variáveis cujos registros superassem 75% do tamanho do histórico dos resultados do sorteio. Com a remoção, a base contabilizou 191 atributos diários meteorológicos.

Uma análise global dos atributos apontou que a ausência sequencial de registros, ainda que menor que 75% do período apurado, poderia fragilizar o futuro modelo. Contando com uma quantidade de atributos suficiente de atributos na base, optou-se pela eliminação daqueles que apresentassem quantidade de registros nulos maior que 150, pois, ainda assim, continuar-se-ia com uma dataset significativo, com registros de 136 atributos válidos.

2.1.3. Histórico de variáveis abstratas/subjetivas

Em busca de uma base diversa, investigou-se variáveis subjetivas e com uma essência mais mística. Reunindo as antigas formas de se prever eventos, ainda presentes na sociedade dos dias atuais, optou-se por variáveis astrológicas que tivessem seu levantamento viável para o estudo. Adotou-se 4 variáveis neste subgrupo: luas, signos, dias da semana e estação do ano nos dias dos sorteios. A fim de não tornar a extração de dados custosa, embora diminuto, o subgrupo se limitou a esses atributos.

Os dias de semana foram extraídos do próprio histórico da Caixa Econômica Federal, e os mesmos foram convertidos para valores numéricos, seguindo o seguinte racional:

DIA DA SEMANA	VALOR NUMÉRICO
Domingo	1
Segunda-feira	2
Terça-feira	3
Quarta-feira	4
Quinta-feira	5
Sexta-feira	6
Sábado	7

Para obtenção das fases da lua relativas aos dias dos sorteios, utilizou-se o site Jorgal, que possui registros das fases lunares no período de 1900 a 2060.⁷ Com a data de alteração de fase, e um tratamento breve no Excel, pode-se preencher todos os dias da base em elaboração. As fases foram transformadas em variáveis numéricas, conforme quadro abaixo.

LUA	VALOR NUMÉRICO
Cheia	1
Minguante	2
Nova	3
Crescente	4

Busca similar foi realizada para o levantamento dos signos vigentes nos dias do sorteio. A disponibilização das datas e períodos de vigência de cada signo foi obtida junto ao site Personare e, também, com um leve tratamento no Excel pode-se alcançar os signos das datas da base dos sorteios.⁸ As datas de início de cada vigência consideradas para a base estão descritas abaixo. Cada signo foi representado por um valor numérico, seguindo racional descrito.

SIGNO	DATA DE INÍCIO	VALOR NUMÉRICO
Peixes	19/fev	1
Áries	21/mar	2
Touro	20/abr	3
Gêmeos	21/mai	4
Cancer	22/jun	5
Leão	23/jul	6
Virgem	23/ago	7
Libra	23/set	8

Escorpião	23/out	9
Sagitário	22/nov	10
Capricórnio	22/dez	11
Aquário	20/jan	12

As estações do ano foram retiradas de site da Fiocruz, que disponibiliza as datas de início das mesmas.⁹ Igualmente, um valor numérico foi atribuído para registro na base, seguindo o mesmo racional das demais variáveis subjetivas.

ESTAÇÃO	DATA DE INÍCIO	VALOR NUMÉRICO
Verão	21/dez	1
Outono	21/mar	2
Inverno	21/jun	3
Primavera	23/set	4

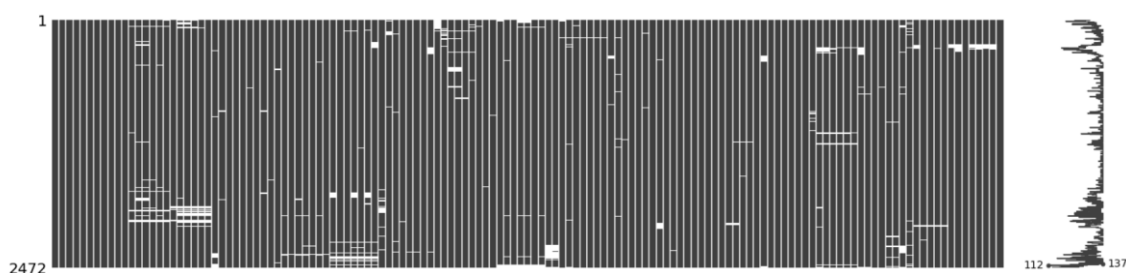
Ao final do processo de localização da informação na web, extração e seleção dos atributos que farão parte do estudo, o banco de dados consolidado contou com 2.472 registros de sorteios entre o período de 11 de março de 1996 e 16 de abril de 2022, e 195 atributos.

Os arquivos com os dados desta subseção e seu tratamento em Excel encontram-se nos anexos deste estudo.

2.2. Análise exploratória e Pré-tratamento em Python

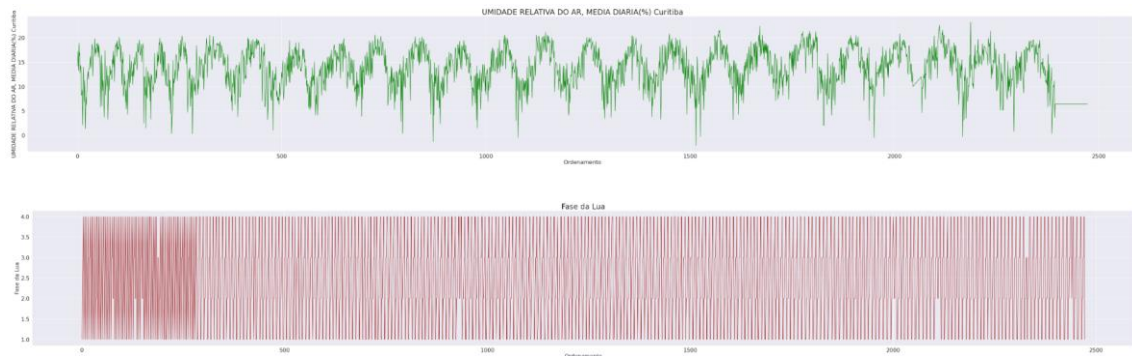
O trabalho e o modelo proposto foram desenvolvidos utilizando a ferramenta Python no ambiente do Google Colab. O arquivo .xlsx da base de dados foi inserido no ambiente virtual e importado para o notebook.

Em um primeiro momento foi realizada uma análise exploratória da base de dados gerada, para maior compreensão do seu conteúdo. Foi montada uma matriz para visualização rápida de valores ausentes e sua distribuição pelos atributos:

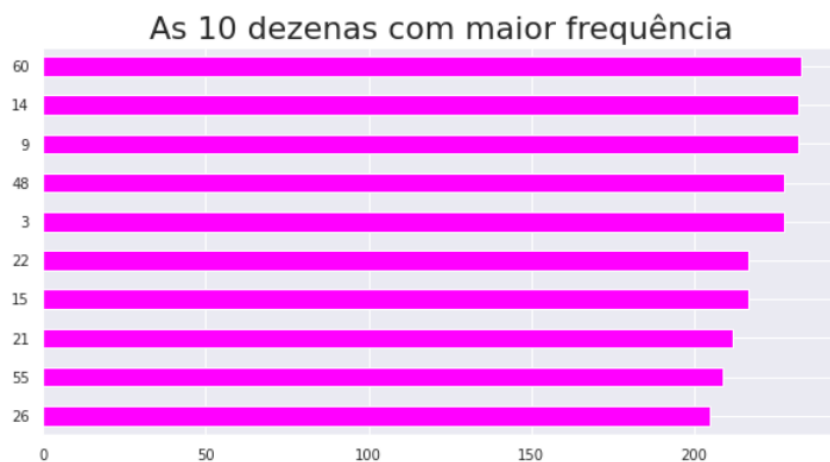


Foram elaborados gráficos para visualização do comportamento de alguns atributos da base, evidenciando comportamentos sazonais em alguns atributos, ao passo que outros não puderam ter qualquer padrão identificável.

A umidade relativa do ar de Curitiba, por exemplo, aponta comportamento cíclico compatível com as estações do ano, de forma similar às fases da lua, cuja alteração no comportamento observado no quarto inicial do gráfico seja relativa à alteração na frequência dos sorteios da base. Ambos podem ser visualizados abaixo, respectivamente.



Um gráfico com a representação das 10 dezenas mais sorteadas na base também foi elaborado, mostrando que dezenas como 60, 14, 9 e 48 saem com maior frequência.



O notebook descritivo pode ser acessado [aqui](#).

2.3. Tratamento de variáveis categóricas - dummies

Após a análise descritiva, avançou-se para a elaboração do modelo preditivo propriamente dito. As bases .xlsx foram carregadas para novo ambiente do Google Colab, e realizou-se a conferência de sua correta transferência. A última coluna da base foi removida por se tratar de uma coluna com o resultado da multiplicação das dezenas, que não será utilizada no modelo.

O trabalho de data mining começou com o tratamento das variáveis categóricas, relativas à lua, signo, estação e dia da semana dos sorteios. Elas foram transformadas em variáveis dummies, de modo que o seu valor numérico atribuído no preenchimento não configurasse qualquer peso equivocado ao atributo, o que poderia interferir nas relações do modelo posteriormente.

2.4. Tratamento de valores ausentes

Os valores ausentes foram tratados com o comando de impute de KNN (k-nearest neighbors) da biblioteca sklearn. Foi utilizado $k = 5$, de modo que os valores ausentes acompanhassem de alguma forma o comportamento dos 5 registros mais próximos.

2.5. Separação na base de treino e teste

O dataset foi, então, separado em base de treino e teste, na proporção de 75%-25%, através do comando `train_test_split`, também do `sklearn`. A separação foi aleatória, sem que houvesse qualquer consideração temporal dos dados.

2.6. Tratamento dos dados de ENTRADA – normalização dos dados numéricos

As variáveis numéricas com registros climáticos da base de treino foram normalizadas, também com a utilização da biblioteca do `sklearn`, e a métrica foi aplicada na base de teste.

A normalização de dados numéricos de um conjunto é benéfica pois reduz a redundância dos dados e nivela os pesos quando atributos de uma mesma base tem dimensões e ordens de grandeza distintas.

2.7. Tratamento dos dados de SAÍDA - transformação das dezenas sorteadas em um array

Os dados de saída continham 6 atributos, cada um referente à uma dezena sorteada. Uma vez que a ordem de sorteio das mesmas não altere o resultado e, buscando uma saída única para cada evento, transformou-se os 6 dados de saída de cada evento em um array com 6 registros.

Foi criada uma função que transformasse essa saída array de 6 elementos em uma saída binária com 60 valores, sendo o valor 1 atribuído à localização correspondente na sequência do novo array. Ex: para um sorteio cujo resultado foi 1, 3, 5, 7, 8 e 9, o novo array ficou: [1, 0, 1, 0, 1, 0, 1, 1, 1, 0, ... + 50 zeros]. Esta transformação permitirá a elaboração de uma rede neural com 60 neurônios na saída.

2.8. Rede neural proposta – Rede Neural Multi-label

Seguindo a lógica de elaboração, optou-se pela construção de uma rede neural multi-label, onde cada uma das dezenas possíveis foi transformada em uma classe de saída, representada por 60 neurônios.

A entrada da rede teve o tamanho idêntico ao número de atributos da base final de entrada – `x_train` e `x_test` – equivalente aos 153 atributos. Foi utilizada a biblioteca `tensorflow`, `keras` para a construção do modelo. Foram criadas 6 camadas densas, com, respectivamente, 80-200-300-200-80-60 neurônios, totalizando 169.960 parâmetros. Optou-se por uma função sigmóide na ativação da 1ª camada e as camadas intermediárias foram testadas com sigmóides e tangente hiperbólica. A camada de saída finalizou com uma função softmax. A arquitetura final da rede proposta está representada na figura a seguir.

```

Model: "sequential"
-----
Layer (type)                 Output Shape                 Param #
-----
dense (Dense)                (None, 80)                  12320
dense_1 (Dense)              (None, 200)                 16200
dense_2 (Dense)              (None, 300)                 60300
dense_3 (Dense)              (None, 200)                 60200
dense_4 (Dense)              (None, 80)                  16080
dense_5 (Dense)              (None, 60)                  4860
-----
Total params: 169,960
Trainable params: 169,960
Non-trainable params: 0
-----

```

2.9. Variáveis do modelo de rede treinado

A rede foi treinada com diferentes modelos, buscando-se aquele que melhor resultados apresentasse. Foram alteradas as quantidades de neurônios nas camadas densas (exceto a camada de saída, que permaneceu com 60 neurônios em todos os modelos), de épocas de treinamento e batch-size. Diferentes funções de ativação foram testadas nas camadas, porém sempre iniciando com uma sigmóide e terminando com uma softmax. Otimizadores e custo de perda do modelo também foram alterados, de modo a se explorar a existência de uma melhor configuração. Os resultados foram organizados apontando as variações utilizadas e os resultados de acurácia de cada modelo.

2.10. Exibição da relação realizado X previsão do modelo: exemplo

Uma vez tendo se treinado o modelo, e obtido seus resultados, foi construída uma saída no notebook para visualização dos valores previstos e realizados de um sorteio qualquer na base de teste. A primeira coluna exibe o resultado do sorteio em questão. Na segunda coluna são apresentadas as seis dezenas cujas saídas da rede treinada apresentaram o maior valor. E na terceira coluna, o valor dessas saídas previstas pelo modelo.

	Dezenas Sorteadas	Dezenas Previstas	Probabilidade da Previsão
0	16	30	0.034367
1	35	10	0.033657
2	36	42	0.032398
3	41	41	0.031336
4	52	27	0.029088
5	54	16	0.028012

3. Resultados

Os resultados obtidos com as configurações de rede testadas foram compilados em tabela e estão representados abaixo. Valores de test score e acurácia estão truncados. Os notebooks de cada modelo podem ser acessados nos arquivos deste trabalho.

Todos os modelos tiveram entrada de 153 neurônios (quantidade de atributos das bases x_train e x_test) e saída softmax de 60 neurônios (cada um representando a probabilidade de sorteio de cada uma das 60 dezenas possíveis. Do tamanho do y_train e y_test).										
Modelo	Camadas Intermediárias	Função de Ativação	Otimizador	Loss	Métrica	Batch size	Épocas	Test Score	Test Accuracy	
01	80/200/300/200/80	Sigmóide e tanh	SGD	categorical_crossentropy	accuracy	100	100	25.69	0.003	
02		Sigmóide						26.09	0.0	
03						200	300	26.59	0.0	
04		26.27						0.067		
05		26.45						0.0		
06	80/500/300/500/100	Sigmóide				26.05		0.067		
07		Sigmóide e tanh				25.96		0.0		
08	80/200/300/200/80	Sigmóide		ADAM		binary_crossentropy	100	100	0.35	0.0
09	80/500/300/500/100						300	0.33	0.008	
10								0.36	0.053	
11			0.326					0.085		
12			1.000	0.378			0.042			
13	10.000	1.389	0.019							
14				1.390		0.017				
15							categorical_crossentropy	41.43	0.027	

Nenhuma das configurações apresentou resultado satisfatório, e a capacidade preditiva do modelo se mostrou semelhante às probabilidades de ocorrência de um evento aleatório. Em todos os modelos a acurácia tendeu a zero, os custos de treino e validação não apontaram ganhos que pudessem indicar qualquer previsibilidade além das previsões probabilísticas de eventos aleatórios.

4. Conclusões

São inúmeras as possibilidades de estudo de eventos aleatórios, e o trabalho realizado é apenas uma forma de realizar previsões do evento aqui apresentado. Outras abordagens podem ser feitas, outras métricas e muitas outras condições podem ser levantadas e aprofundadas. A exaustão dessas perspectivas é incalculável.

Embora haja uma quantidade enorme de dados disponíveis na web e a capacidade de aprendizado dos algoritmos de machine learning seja surpreendente, a dificuldade de seleção dos dados verdadeiramente correlacionados ao evento pode ser parte do grande desafio. A fidelidade dos dados e obtenção de bases robustas também elevam os obstáculos. A quantidade de variáveis associadas pode ser imensa, e seu levantamento impraticável. Ainda que haja um algoritmo capaz de realizar todas as previsões, os eventos aleatórios ainda seriam uma forma de apresentar as incertezas e riscos inerentes ao limitado conjunto de informações que temos sob nosso conhecimento prévio.

O espaço amostral faz referência aos diferentes possíveis resultados de um processo aleatório. Em casos simples, o espaço consiste em alguns pontos, ao passo que para processos complexos podem tomar uma dimensão contínua, aproximando-se ao espaço exatamente igual ao que vivemos. E assim, os fatores que influenciam os eventos aleatórios podem ser (e geralmente o são) tão complexos que qualquer tentativa de previsão se torne tão vulnerável a influências imprevisíveis e incontroláveis que simples palpites bem informados podem chegar às mesmas virtudes. Muitas das vezes, acabam sendo tão assertivos quando qualquer chute de um desavisado.

Ainda há muito o que se desenvolver quanto ao registro fiel de dados de acontecimentos, seu armazenamento e compartilhamento. E muito a se aprender das relações no mundo que nos cerca para que se possa compreender melhor a previsibilidade do chamado aleatório. Por ora, a previsão das dezenas do jogo de azar segue sendo calcada em intuições, achismos, vontade dos deuses ou como se quiser denominar a ciência do acaso.

Referências bibliográficas

1. Mlodinow, L. **O andar do bêbado: como o acaso determina nossas vidas**. 1.ed. Rio de Janeiro. Editora Schwarcz. 2009.
2. Agrawal, A; Gans, J.; Goldfarb, A. **Máquinas preditivas: a simples economia da inteligência artificial**. 1 ed. Rio de Janeiro. Editora Alta Books. 2018.
3. Domingos, P. **O Algoritmo Mestre: Como a Busca Pelo Algoritmo de Machine Learning Definitivo Recriará Nosso Mundo**. 1.ed. São Paulo. Novatec Editora. 2017.
4. SAS. **Machine Learning: O que é e qual sua importância?** Disponível em: https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Acesso em: 30/05/2022.
5. Caixa Econômica Federal, 2022. **Resultados sorteio Mega Sena**. Disponível em: <https://loterias.caixa.gov.br/Paginas/Download-Resultados.aspx>. Acesso em: 24/05/2022.
6. Instituto Nacional de Meteorologia. Ministério da Agricultura, Pecuária e Abastecimento, 2022. **Banco de dados meteorológicos**. Disponível em: <https://bdmep.inmet.gov.br/>. Acesso em: 30/05/2022.
7. Jorgal, 2022. **Fases da Lua do Ano 1900 Até 2060**. Disponível em: <https://www.jorgal.com.br/fases-da-lua/>. Acesso em: 25/05/2022.
8. Personare, 2022. **Data dos signos**. Disponível em: <https://www.personare.com.br/signosPersonare>. Acesso em: 25/05/2022.
9. Fundação Oswaldo Cruz, 2022. **Estações do ano**. Disponível em: <https://www.fiocruz.br/biosseguranca/Bis/infantil/estacoes-ano.html>. Acesso em: 25/05/2022.

Matrícula: 202.190.064

Pontifícia Universidade Católica do Rio de Janeiro

Curso de Pós Graduação Business Intelligence Master