Análise Exploratória de Datasets Públicos Para Manutenção Preditiva

Pedro Inácio de Oliveira Filho, Eduardo Camilo Inacio

[1]Centro Universitário SENAI/SC - Campus Florianópolis , Florianópolis, SC, Brasil

pedroinaciooliveira@hotmail.com, eduardo.inacio@edu.sc.senai.br

Resumo. *O constante desenvolvimento tecnológico de algoritmos de Machine Learning voltados à manutenção preditiva, viabilizam a realização de estudos e pesquisas envolvendo o tema proposto. A análise exploratória dos dados e a estatística descritiva são de suma importância na construção destas pesquisas, haja vista que, possibilita melhor compreensão e clareza no aprendizado e nas pesquisas relacionadas à manutenção preditiva. O artigo realiza uma análise exploratória de Datasets públicos para manutenção preditiva, no intuito de reunir resultados obtidos por meio do estudo exploratório dos dados e implementar análises descritivas preliminares nos dados encontrados nestes Datasets. Objetiva-se, ainda, destacar a importância deste processo de análise estatística na compreensão introdutória ao processo de aprendizado para ciência de dados*
.

1. Introdução

Este artigo artigo tem como foco precípuo, uma análise realizada por meio da exploração de dados, utilizando como ferramenta principal, o Pandas (biblioteca em Python que contém estruturas de dados e ferramentas para manipulação de dados). Para isso, foram realizadas consultas e estudos em datasets públicos, visando esclarecer dúvidas sobre as principais técnicas de análise exploratória de dados e aplicação dessas técnicas no desenvolvimento deste artigo.

A análise estatística exploratória contribui diretamente ao processo de estudo dos dados obtidos, havendo assim, melhor compreensão do que será trabalhado na pesquisa inicial. A análise descritiva de dados, limita-se a calcular algumas medidas de posição e variabilidade, como a média e variância, por exemplo. Em outra corrente mais moderna, utiliza-se, principalmente, técnicas gráficas, em oposição a resumos numéricos. Isso não significa que sumários não devam ser obtidos, mas uma análise exploratória de dados não deve se limitar a calcular tais medidas [Morettin and Bussab 2017]. Foram utilizados métodos gráficos obtidos da documentação do Pandas para plotar os dados originais, no intuito de buscar informações claras e precisas sobre o conteúdo desses dados.

2. Fundamentação Teórica

A construção teórica desta pesquisa, permeia estudos introdutórios da mineração de dados como: pré-processamento de dados, visualização, análise e associação e modelagem preditiva. Este artigo delimita-se, prioritariamente, na análise exploratória como objeto principal de desenvolvimento do estudo.

A mineração de dados surge a partir do avanço rápido das tecnologias de coleta e armazenamento. Com o crescente número de dados coletados nas organizações, fez se necessária a criação de ferramentas e técnicas de exploração dos dados, para que fosse possível analisar, de forma mais adequada, as informações encontradas nos conjuntos de dados obtidos. A mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados. Ela também abriu oportunidades interessantes para se explorar e analisar novos tipos de dados e para analisar tipos antigos de novas maneiras [Tan et al. 2009].

Com a disponibilidade de capacidade computacional e expressivos softwares de análise de dados, a análise exploratória de dados evoluiu muito além de seu escopo original. As principais características dessa modalidade tem sido o rápido desenvolvimento de novas tecnologias, o acesso a dados maiores e em maior quantidade e o uso de análises quantitativas em diversas modalidades [Bruce and Bruce 2019].

Dentro das análises estatísticas, podemos destacar como objeto de estudo, a análise exploratória, que representa a técnica de explorar previamente as informações de um conjunto de dados para que, posteriormente, ele seja utilizado em uma aplicação de negócio, estatística ou de aprendizado de máquina. A mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de mineração de dados são organizadas para agir sobre grandes bancos de dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados. Contudo, a mineração de dados tem sido usada para melhorar sistemas de recuperação de informações [Tan et al. 2009].

Os dados podem ser obtidos de diversas fontes, por textos, imagens, vídeos, medições etc. Grande parte desses não são dados estruturados, portanto, para aplicarmos conceitos estatísticos, é necessário manipular e processá-los para que tomem uma forma estruturada. Dentre os dados estruturados, podemos classificar duas formas: a forma contínua, como velocidade do vento ou tempo de duração, e a forma discreta, como a contagem de ocorrências de um evento. Os dados categóricos assumem apenas um conjunto fixo de valores, enquanto os dados binários são do tipo categórico e assumem apenas um de dois valores, como 0 ou 1, sim ou não,verdadeiro ou falso. os dados ordinais também são um tipo de dado categórico, no qual as categorias são ordenadas, como por exemplo, uma classificação numérica. Para fins de análise de dados e modelagem preditiva, o tipo de dados é importante porque ajuda a determinar o tipo de exposição visual, análise de dados ou modelo estatístico. Além disso, o tipo de dados para uma variável, determina como o software processará os cálculos para aquela variável [Bruce and Bruce 2019].

No capitulo seguinte, far-se-á uma abordagem teórica sobre estatística, especificando técnicas, medidas e modelos estatísticos que são de suma importância para introdução à mineração de dados.

2.1. Medidas Estatísticas

A exploração dos dados pode ajudar na escolha das técnicas adequadas de pré-processamento e análise dos dados. Ela pode até mesmo abordar algumas das questões que são geralmente respondidas pela mineração de dados. Por exemplo, os padrões às vezes podem ser encontrados através da inspeção visual dos dados. Além disso, algumas das técnicas usadas na exploração dos dados, como a visualização, podem ser usadas para

se entender e interpretar os resultados da mineração de dados [Tan et al. 2009].

2.1.1. Medidas de Localização: média e mediana

No uso de dados contínuos, as estatísticas de resumo frequentemente utilizadas são a média e a mediana. Ambas são medidas da localização de um conjunto de valores. A mediana é o valor intermediário se houver um número ímpar de números e a média dos dois valores do meio se o número de valores for par. Embora a média seja interpretada às vezes como o meio de um conjunto de valores, isto só está correto se os valores estiverem distribuídos de uma forma simétrica. Se a distribuição de valores for irregular, então a mediana é um indicador melhor do meio. Além disso, a média é sensível à presença de externos. Para dados com externos, a mediana fornece novamente uma avaliação mais robusta do meio de um conjunto de valores [Tan et al. 2009].

2.1.2. Medidas de Dispersão: Faixa e variância

No uso de dados contínuos utiliza-se, normalmente, a estatística de resumo, que serve para aplicar a medição da dispersão ou dispersão de um conjunto de valores. Tais medidas indicam se os valores dos atributos estão muito dispersos ou se estão relativamente concentrados em um único ponto como a média. A medida mais simples da dispersão é a faixa (variação), a qual, dado um atributo x com um conjunto de valores $[x_1, ..., x_m]$, é definida como [Tan et al. 2009].

$$faixa(x) = máx.(x) - min.(x) = x_{(m)} - x_{(1)} \tag{1}$$

Num conjunto de dados, a variância é uma medida de dispersão que especifica a distância que cada valor está em relação à média (ponto central). Quanto mais próximo da média, menor é a variância, quanto mais distante da média, maior será a variância.

Embora a faixa identifique a dispersão máxima, ela pode levar a enganos se a maioria dos valores estiver concentrada em uma faixa estreita de valores, mas também, se houver um número relativamente pequeno de valores mais extremos. A variância dos valores de um atributo x geralmente é descrita como s_x^2 e está definida a seguir. O desvio padrão, que é a raiz quadrada da variância, é escrito como s_x e possui as mesmas unidades de x.

$$variância(x) = S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m} = (xi - \overline{x})^2 \tag{2}$$

A média pode ser distorcida pelos externos e, já que a variância é calculada usando a média, também é sensível aos externos. De fato, a variância é especialmente sensível a externos, já que usa a diferença quadrada entre a média e outros valores. Como consequência, avaliações mais robustas da dispersão de um conjunto de valores são usadas frequentemente [Tan et al. 2009].

2.2. Variáveis

A utilização da estatística como ferramenta de análise de dados é essencial, e para melhor definição, separa-se em dois ramos diferentes de pesquisa: o Qualitativo e o Quantitativo.

Estes meios de demostrar dados e adquirir informações ambos com diferentes métodos de coleta e amostra, resultam em variáveis que podem ser classificadas de acordo com o teor de sua amostra. Algumas variáveis como sexo, educação, estado civil, apresentam como possíveis realizações uma qualidade (ou atributo) do indivíduo pesquisado, ao passo que outras, como o número de filhos, salário, idade,apresentam como possíveis realizações números resultantes de uma contagem ou mensuração. As variáveis do primeiro tipo são chamadas qualitativas, e as do segundo tipo, quantitativas [Morettin and Bussab 2017].

Variáveis com dados de medição ou contagem podem ter milhares de valores diferentes. Um passo fundamental na exploração de seus dados e definir um "valor típico"para cada característica (variável): uma estimativa de onde a maioria dos dados está localizada (ou seja, usa tendência central) [Bruce and Bruce 2019].

Ainda sobre as variáveis qualitativas, existem distinções entre qualitativa nominal, para qual não existe ordenação nas possíveis realizações, e variável qualitativa ordinal, para a qual existe uma ordem nos seu resultados. Quanto às variáveis quantitativas, elas podem ser classificadas como quantitativas discretas: cujos possíveis valores formam um conjunto finito ou enumerável de números, e quantitativas contínuas: cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração [Morettin and Bussab 2017].

2.3. Distribuições de Frequências

Ao estudarmos uma variável, a distribuição de frequências advém especificamente da necessidade de conhecermos a atuação dessa variável, analisando a ocorrência de suas possíveis movimentações.

Dependendo do volume de dados, torna-se difícil ou impraticável tirar conclusões a respeito do comportamento das variáveis e, em particular, de variáveis quantitativas. Pode-se, no entanto, colocar os dados brutos de cada uma das variáveis quantitativas em uma ordem crescente ou decrescente, denominado rol. A visualização de algum padrão ou comportamento continua sendo de difícil observação ou até mesmo cansativa, mas torna-se rápido identificar maiores e menores valores ou concentrações de valores no caso de variáveis quantitativas. Estes números (menor e maior valor observado) servem de ponto de partida para a construção de tabelas para estas variáveis. Vale destacar que para as variáveis qualitativas, pode- se também construir um rol em ordem temporal ou alfabética, por exemplo. É a diferença entre o menor e maior valor observado da variável X, denominada amplitude total ($AT = x_{max}–x_{min}$), que definirá a construção de uma distribuição de frequência pontual ou em classes. O ideal é que uma distribuição de frequência resuma os dados em um número de linhas que varie de 5 a 10 [Guedes et al. 2005].

O modo é o valor mais frequente da distribuição de frequências. Para variáveis discretas, o modo é definido pela verificação da tabela de frequências, sendo atribuído ao tipo que tem maior frequência. Dado um conjunto não ordenado de valores categorizados, não há muito a ser feito para caracterizar melhor os valores, além de calcular a frequência na qual cada valor ocorre em um determinado conjunto de dados. O modo de um atributo categorizado é o valor que possui a frequência mais alta. Para dados contínuos, o modo, conforme definido correntemente, muitas vezes não é útil porque um único valor pode não ocorrer mais que uma vez, além disso, se um valor único for usado para indicar a falta de um valor, então este valor muitas vezes aparecerá como o modo [Tan et al. 2009].

O resumo de dados por meio de tabelas de frequências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados [Morettin and Bussab 2017].

3. Procedimentos Metodológicos

Após discorrermos sobre as principais etapas da Estatística Descritiva, foi feita uma busca por datasets para serem utilizados como material de estudo, a fim de validar a proposta de análise exploratória em datasets públicos para manutenção preditiva.

Para gerar as estatísticas descritivas utilizamos o método ”describe()”. As estatísticas deste método incluem aquelas que resumem a tendência central, dispersão e forma da distribuição de um conjunto de dados, excluindo valores nulos. Também analisa séries numéricas e de objetos, bem como conjuntos de colunas de tipos de dados mistos. A saída irá variar dependendo do que é fornecido. Foram selecionados três datasets que serão, a seguir, analisados individualmente, a fim de descrever suas estatísticas.

3.1. Data Set 01 - Elevator Predictive Maintenance

O primeiro dataset a ser explorado, ”Elevator Predictive Maintenance Dataset”, é de domínio publico e foi acessado na plataforma Kaggle. Este dataset possui um conjunto de dados de manutenção preditiva para previsão de falhas de elevadores do Huawei German Research Center. O dataset possui dados de sensores IoT para manutenção preditiva no setor de elevadores. Sua principal utilidade será na manutenção preditiva de portas de elevadores, a fim de reduzir paradas não planejadas e maximizar o ciclo de vida dos equipamentos.

O dataset contém dados de operação na forma de séries temporais, amostradas a 4 Hz no uso de um elevador em horário de pico noturno, em um prédio (entre 16:30 e 23:30). Na porta do elevador, observamos as seguintes características de funcionamento: o sistema de sensores eletromecânicos (Sensor de rolamento de esferas da porta), os dados climáticos (umidade) e físicos (Vibração). Com o objetivo de obter o valor absoluto da vibração.

Tabela 1. Dados estatísticos do dataset: Elevator Predictive Maintenance

	count	mean	std	min	25%	50%	75%	max
ID	112001.0	56001.000000	32332.048087	1.000000	28001.000000	56001.000000	84001.000000	112001.000000
revolutions	112001.0	46.275195	19.042179	16.933000	29.651000	43.348000	63.997000	93.744000
humidity	112001.0	74.224140	0.684711	72.399000	73.914000	74.212000	74.731000	75.400000
vibration	109563.0	28.340276	24.292500	2.000000	8.000000	21.280000	39.210000	100.000000
x1	112001.0	120.499335	18.984921	90.132000	103.850000	117.640000	138.119000	167.743000
x2	112001.0	-27.948945	19.123796	-56.353000	-44.548000	-31.443000	-10.012000	19.745000
x3	112001.0	0.623759	0.258677	0.231328	0.399615	0.580561	0.865330	1.266828
x4	112001.0	2503.994994	1874.972912	286.726489	879.181801	1879.049104	4095.616009	8787.937536
x5	112001.0	5509.691804	101.395621	5241.615201	5463.279396	5507.420944	5584.722361	5685.160000

Conforme mostra a tabela, ao aplicar o método ”describe()”obteve-se os seguintes parâmetros estatísticos: A contagem foi de 01 a 112001, sendo que, a quantidade de registros de vibração ficou em 109563 registros. A maior média registrada entre so sensores foi 5509 no sensor x5. O maior valor registrado para o desvio padrão foi 1874 no sensor x4. Notou-se um aumento considerável na umidade registrada, variando entre o mínimo de 72,39 e o máximo de 75,40. A média de vibração entre o mínimo 2 e máximo

100, ficou registrada em 28,3. Os sensores de 1 a 5 registraram números totalmente independentes, sendo que os valores máximos registrados para cada sensor foi: sensor x1 registrou máx. 167,7. Sensor x2 registrou máx. 19,74. Sensor x3 registrou máx. 1,26. Sensor x4 registrou máx. 8787 e o sensor x5 registrou 5685. Portanto, percebe-se, com as revoluções registradas, que houve aumento gradativo de umidade e vibração conforme os registros foram sendo captados, sendo que o terceiro quantil registra a maior incidência de vibração e umidade.

3.2. Data Set 02 - Synchronous Machine

Este Data Set público foi acessado na UCI Machine Learning Repository. Conforme as informações obtidas nas especificações, os dados da máquina síncrona foram captados em tempo real, no ambiente operacional experimental. As colunas demonstram os seguintes dados: ”Iy”(Corrente de carga); ”PF”(Fator de potência); ”e”(Erro de fator de potência); ”dIf”(Mudança da corrente de excitação da máquina síncrona); ”If”(Corrente de excitação da máquina síncrona).

Motores síncronos (SMs) são motores CA com velocidade constante. Um conjunto de dados SM é obtido a partir de um conjunto experimental real. A tarefa descrita nas informações obtidas do Dataset é criar modelos fortes para estimar a corrente de excitação de SM.

Aplicou-se o método ”describe()”para obter os valores estatísticos e foi realizada também a transposição da tabela, portanto, foram retornados os seguintes dados:

Tabela 2. Estatísticas do dataset: Synchronous Machine

	count	mean	std	min	25%	50%	75%	max
Iy	557.0	4.499820	0.896024	3.000	3.700	4.500	5.300	6.000
PF	557.0	0.825296	0.103925	0.650	0.740	0.820	0.920	1.000
e	557.0	0.174704	0.103925	0.000	0.080	0.180	0.260	0.350
dIf	557.0	0.350659	0.180566	0.037	0.189	0.345	0.486	0.769
If	557.0	1.530659	0.180566	1.217	1.369	1.525	1.666	1.949

Após transpostas, as colunas serão tratadas como linhas, portanto, na primeira linha, denominada ”Iy”, o contador (count) revela a incidência de 557 testes (valor obtido para todas demais linhas da coluna ”count”, sendo que, a média registrada para a corrente de carga ficou em 4,49, com desvio padrão de 0,89, mínima 3,0 e máxima 6,0

Na linha ”PF”, a média registrada para o fator de potência, foi 0,82, o desvio padrão 0,10 e o máximo registrado foi 1.

A linha ”e”registrou uma média 0,17 de erro no fator de potência, atingindo o máximo 0,35. Conclui-se que a maior quantidade de erros foram registrados no segundo quartil (Q2)

Na linha ”dIf”foi registrada uma média de 0,35 na mudança de corrente de excitação da máquina. A linha ”If”, retorna a corrente de excitação máxima registrada em 1,49. Portanto, nota-se que o desvio padrão repete-se nas linhas de fator de potência e

erro de fator de potência (”PF”e ”e”), assim como nas linhas que representam a corrente de excitação e máxima corrente de excitação (”dIf”e ”If”).

3.3. Data Set 03 - Machine Predictive Maintenance Classification

O terceiro dataset a ser utilizado, ”Machine Predictive Maintenance Classification”, acessado através da plataforma Kaggle, fornece um conjunto de dados sintéticos que refletem a manutenção preditiva real encontrada no setor. O conjunto de dados consiste em 10.000 registros com 11 colunas. Para visualização da tabela, foram removidos dados categóricos no formato de texto. Estes dados foram desconsiderados na análise estatística, portanto, foram mantidas 7 colunas para a criação da tabela a seguir:

Tabela 3. Estatísticas do Dataset: Machine Predictive Maintenance Classification

	count	mean	std	min	25%	50%	75%	max
UDI	10000.0	5000.50000	2886.895680	1.0	2500.75	5000.5	7500.25	10000.0
Air temperature [K]	10000.0	300.00493	2.000259	295.3	298.30	300.1	301.50	304.5
Process temperature [K]	10000.0	310.00556	1.483734	305.7	308.80	310.1	311.10	313.8
Rotational speed [rpm]	10000.0	1538.77610	179.284096	1168.0	1423.00	1503.0	1612.00	2886.0
Torque [Nm]	10000.0	39.98691	9.968934	3.8	33.20	40.1	46.80	76.6
Tool wear [min]	10000.0	107.95100	63.654147	0.0	53.00	108.0	162.00	253.0
Target	10000.0	0.03390	0.180981	0.0	0.00	0.0	0.00	1.0

Para a criação da tabela de dados estatísticos, foram feitas algumas manipulações para que as colunas (Type, failure Type, Product ID e dtype) fossem removidas. Assim, as demais colunas foram analisadas e apresentaram os seguintes dados estatísticos: a coluna ”UDI”, que define um ID para cada registro do conjunto de dados, apresentou, como esperado, todos os registros necessários, não havendo assim, necessidade de análise estatística para esta coluna.

A coluna ”Air Temperature [K]”, mostra a temperatura do ar em Kelvin, especificamente no ato da falha do equipamento, onde observamos um desvio padrão de 1,48, temperatura mínima registrada de 295,3, chegando à máxima temperatura de 304,5.

A coluna ”Process temperature [K]”representa a temperatura do processo em Kelvin, no ato da falha da máquina, onde observa-se um desvio padrão de 1,48. A média registrada foi 310. Portanto, observa-se que a temperatura do processo registrou média superior à média da temperatura do ar, que ficou em 300.

A coluna ’Rotational [rpm]”mostra a velocidade da rotação por minuto, no ato da falha de máquina. A média rotacional registrada em rpm foi 1538, chegando à máxima de 2886. Podemos observar, então, que a maior variação rotacional foi registrada entre o quartil 3 (Q3) e o valor máximo, aumentando de 1612 rpm a 2886 rpm.

Na coluna ”Torque [Nm]”, temos a representação do torque da máquina em Nm (Newton metros) no momento da falha de máquina, que registrou o torque mínimo de 3,8 Nm e máximo 76,6 Nm, sendo que, a maior variação de torque pode ser observada nos dados dos quartis 1 (Q1), que varia de 3,8 Nm a 33,20 Nm e do quartil 3 (Q3) até valor máximo, que varia entre 46,8 Nm e 76,6 Nm.

A coluna ”Tool wear [min]”mostra o desgaste da máquina no momento em que a falha é registrada. Esta coluna apresentou média de desgaste em 108. Observou-se que o

maior foi registrado entre as colunas que representam o quartil 3 (Q3) e o valor máximo, sendo registrado o aumento de 162 para 253.

Na coluna ”Target”observamos que existem apenas dois tipos de dados: onde registra-se 0 para não falha e 1 para falha. Esta coluna não apresenta dados estatísticos relevantes.

4. Considerações e Trabalhos Furturos

Este trabalho representou um esforço de posicionamento crítico, no que refere-se a iniciação científica, ressaltando a importância que a pesquisa estatística descritiva representa para análise de dados. Constatou-se a importância de utilizar como base de estudos, a aplicação de métodos estatísticos para a obtenção de resultados nos datasets analisados. Foram estudados datasets públicos para manutenção preditiva. As análises estatísticas demonstraram a importância da observação, manipulação e tratamento de dados, além de melhorar a compreensão no estudo de algoritmos de classificação. Portanto, pode-se afirmar que os estudos desenvolvidos na área de estatística e pré-processamento de dados, contribuirão significativamente em outras pesquisas, ampliando as bases e servindo como suporte para demais pesquisas e trabalhos futuros que envolvam Machine Learning e manutenção preditiva.

Referências

Bruce, A. and Bruce, P. (2019). *Estatística Prática para Cientistas de Dados*. Alta Books.

Guedes, T. A., Martins, A. B. T., Acorsi, C. R. L., and Janeiro, V. (2005). Estatística descritiva. *Projeto de ensino aprender fazendo estatística*, pages 1–49.

Morettin, P. A. and Bussab, W. O. (2017). *Estatística básica*. Saraiva Educação SA.

Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao datamining: mineração de dados*. Ciência Moderna.