# Título do Artigo de Conferência:

Avaliação Inicial e Pré-processamento de Grandes Conjuntos de Dados: Revelação de Perspetivas nas Condições Meteorológicas e Consumo de Energia

|  |  |  |
| --- | --- | --- |
| Nuno Gomes  MEIC- Mineração de Dados em Larga Escala  Instituto Superior Engenharia de Lisboa  Lisboa, Portugal  E-mail: A18364@alunos.isel.ipl.pt | Ricardo Ramos  MEIC  Mineração de Dados em Larga Escala  Instituto Superior Engenharia de Lisboa  Lisboa, Portugal E-mail:  46638@alunos.isel.ipl.pt | Rafael Carvalho  MEIC  Mineração de Dados em Larga Escala  Instituto Superior Engenharia de Lisboa  Lisboa, Portugal  E-mail: 47663@alunos.isel.ipl.pt |

Autores:

Resumo:

O objetivo deste trabalho é demonstrar claramente a compreensão do problema de mineração de dados, selecionando e aplicando soluções adequadas ao longo do desenvolvimento do modelo de aprendizagem. Nesta primeira fase, procura-se descrever o problema específico de mineração de dados e enquadrá-lo no contexto geral desta área. Além disto, é caracterizado detalhadamente o conjunto de dados utilizado, explicando os critérios seguidos para sua construção. São indicados os problemas encontrados durante o pré-processamento dos dados e as soluções adotadas.

Introdução:

Perante os grandes conjuntos de dados abarcando condições atmosféricas e consumos energéticos, o objetivo é extrair informações valiosas e padrões úteis. Foi realizada uma avaliação inicial para observar os conjuntos de dados, com o intuito de refinar e gerar hipóteses para a extração de informações e padrões que possam ser conclusivos em relação a um ou mais problemas formulados. Por exemplo, determinar um classificador de zonas residenciais ou industriais com base no consumo energético e/ou gerar um modelo de regressão que permita prever o consumo energético com base nas condições meteorológicas para um determinado código postal, ou mesmo antecipar aumentos de consumo com base tanto nas condições meteorológicas como nos períodos sazonais.

Após esta fase, os dados foram submetidos a um pré-processamento para criar conjuntos de dados que permitam a extração das informações desejadas e treinar modelos para obter classificadores o mais robustos possível. Nesta fase inicial, é imperativo compreender o problema específico a ser abordado, enquadrando-o no contexto mais amplo das tarefas de mineração de dados. Além disto, é essencial caracterizar o conjunto de dados disponível, descrevendo detalhadamente a sua origem, estrutura e qualidade.

Durante o pré-processamento dos dados, são encontrados desafios que exigem soluções adequadas, tais como lidar com valores em falta, normalizar dados e tratar variáveis categóricas. Identificar estes problemas e implementar as soluções mais adequadas é crucial para garantir a qualidade e integridade dos dados analisados.

A aplicação de técnicas de redução de dimensionalidade, abordando o problema da maldição da dimensionalidade e a necessidade de seleção e redução de características, torna-se relevante quando lidamos com conjuntos de dados complexos. Porem, Reduzir a dimensionalidade pode resultar em perda de informações importantes dos dados originais, afetando a precisão do modelo ao aprender padrões. Isso torna a interpretação dos resultados mais desafiadora, já que as relações entre variáveis e padrões nos dados podem-se tornar menos claras. Além disso, a redução de dimensionalidade pode inadvertidamente remover informações úteis, resultando na retenção de ruído nos dados ou na eliminação de características importantes. No entanto, é necessário ter atenção ao risco de exclusão de dados ou enviesamento dos mesmos, o que pode levar a informações distorcidas, mas, a aplicação destas técnicas, quando aplicadas devidamente, permitem simplificar a representação dos dados sem perder informações importantes, facilitando análises subsequentes.

Todo o processo de análise de dados foi realizado utilizando a linguagem de programação R. Foi essencial desenvolver código claro e organizado que execute todas as análises propostas, garantindo a reprodutibilidade e transparência dos resultados obtidos.

Neste processo, também designado de reengenharia de características, são criados os conjuntos de dados fundamentais para o desenvolvimento de modelos de aprendizagem. Ao compreender claramente o problema, caracterizar os dados e aplicar técnicas de pré-processamento adequadas, estamos a preparar o terreno para análises mais avançadas nas fases subsequentes.

Caracterização do conjunto de dados disponível

1. Dados de consumos de energia

Os dados fornecidos são relacionados à energia elétrica, com informações sobre data, hora, código postal e energia ativa em quilowatt-hora (kWh). São provenientes de medições de energia elétrica em diferentes locais, talvez residenciais, comerciais ou industriais, identificados pelos códigos postais.

i) Estrutura dos Dados:

Date/Time (Data/Hora): Representa a data e hora em que a medição foi feita. Parece estar em formato ISO 8601 (com alguns desvios), incluindo o fuso horário.

Date (Data): Data da medição no formato DD/MM/AAAA.

Hour (Hora): Hora da medição.

Zip Code (Código Postal): Código postal do local onde a medição foi feita.

Active Energy (Energia Ativa): Quantidade de energia elétrica ativa consumida, medida em kWh.

ii) Qualidade dos Dados:

Consistência: Os dados parecem consistentes em relação aos tipos de informações fornecidas.

Completude: Não há evidências de valores em falta nas amostras fornecidas.

Precisão: Os valores numéricos, como a energia ativa, parecem ser precisos, com muitos dígitos decimais.

Formato de Data/Hora: Algumas datas e horas parecem estar fora de formato (por exemplo, o campo "Hour" às vezes mostra horas diferentes da hora completa na coluna "Date/Time"). Isto pode precisar ser ajustado para garantir consistência.

Fuso Horário: Há uma mistura de horários com e sem fuso horário. Isto pode precisar de ser padronizado para uma única representação (por exemplo, todas as horas em UTC ou num fuso horário específico).

Observações Adicionais:

Há uma mistura de datas em diferentes formatos (DD/MM/AAAA e YYYY-MM-DD) e horários em diferentes fusos horários. Isto também precisa ser harmonizado para facilitar a análise.

A presença de uma grande quantidade de dígitos decimais na energia ativa pode indicar uma alta precisão nas medições ou a necessidade de arredondamento para simplificar a análise.

Para uma análise mais aprofundada ou para utilização em projetos específicos, pode ser necessário realizar limpezas e transformações adicionais nos dados.

1. Dados de condições climatéricas

Este conjunto de dados retrata uma série temporal de observações meteorológicas, provenientes de uma estação meteorológica local ou de uma fonte semelhante que regista informações climáticas em Lisboa.

i) Estrutura dos Dados:

Cada linha representa uma observação de hora a hora, com os seguintes parâmetros/colunas:

name: Nome da localidade (constantemente "Lisbon" neste conjunto).

datetime: Data e hora da observação no formato "AAAA-MM-DDTHH:MM:SS".

temp: Temperatura em graus Celsius.

feelslike: Sensação térmica em graus Celsius.

dew: Ponto de orvalho em graus Celsius.

humidity: Umidade relativa em percentual.

precip: Precipitação em milímetros.

precipprob: Probabilidade de precipitação em percentual.

preciptype: Tipo de precipitação (por exemplo, "rain" para chuva).

snow: Quantidade de neve em milímetros.

snowdepth: Profundidade da neve em milímetros.

windgust: Rajada máxima do vento em quilómetros por hora.

windspeed: Velocidade do vento em quilómetros por hora.

winddir: Direção do vento em graus.

sealevelpressure: Pressão atmosférica ao nível do mar em milibares.

cloudcover: Cobertura de nuvens em percentual.

visibility: Visibilidade em quilómetros.

solarradiation: Radiação solar em Watts por metro quadrado.

solarenergy: Energia solar em Joules por metro quadrado.

uvindex: Índice UV.

severerisk: Risco severo.

conditions: Condições meteorológicas resumidas.

icon: Ícone representando as condições meteorológicas.

stations: Lista de estações meteorológicas.

ii) Qualidade dos Dados:

Qualidade dos dados: Para avaliar a qualidade dos dados, seria necessário analisar vários aspectos, como a consistência das observações, a ausência de dados faltantes ou inconsistentes, e a comparação com outras fontes confiáveis de dados meteorológicos para verificar a precisão das medições.

Pontos de atenção:

Pode ser necessário realizar um pré-processamento nos dados, como tratamento de valores ausentes, correção de erros de digitação e normalização de unidades.

Além disso, é importante considerar a estação do ano e as condições locais ao interpretar os dados, já que o clima pode variar significativamente ao longo do ano em Lisboa.

Estas informações são essenciais para entender e utilizar eficazmente o conjunto de dados meteorológicos disponíveis.