Análise de dados Exploratória: aplicando métodos para padronização e normalização dos dados

Sumário

Motivações

Está pesquisa tem como propósito aplicar métodos voltados para **análise exploratória** dos dados que por sua não tem como objetivo entregar insghts para às partes interessadas, uma vez que este tipo de análise é atribuído para a **análise de dados explanatória.** Aqui vamos discutir sobre técnicas e métodos aplicados para um melhor entendimento dos dados, entender como outros analistas decidem aplicar métodos X ao invés de Y e assim por diante, vale lembrar que todos os artigos ou referências usadas serão deixas na página **notion.**

Gerenciador de pacotes e módulos utilizados na pesquisa

Acredito que não seria nem um pouco encorajador ver uma vasta lista de módulos ou até mesmo softwares que no fim nem se quer foram usados, portanto prefiro fugir do complexismo para de algum modo gourmetizar meu trabalho. Toda a pesquisa será feita utilizando Python como ferramenta para analisar os dados, tanto na questão estatística quanto visualização dos dados.

- Gerenciador: Conda
- Pacotes: pyplotly, seaborn, scipy, sklearnig e pandas

Especificações do hardware usado

- Modelo: Dell Vostro 3501
- Sistema operacional: Ubuntu 25.04
- Memória principal (RAM): 12 GB
- Memória interna: 256 GB

Convenções utilizadas

Logo a baixo você poderá ver algumas convenções utilizadas para a criação deste documento, vale ressaltar que apesar daqui não se tratar de um “guia faça você mesmo” que passa explicar um passo a passo das instruções criadas, ainda sim gostaria de utilizar uma fonte especifica para os notebooks aqui mostrados.

`Liberation Mono`

Ao invés de utilizar imagens de cada notebook escrito, optei por utilizar está fonte para cada código feito para está pesquisa.

Contato

Caso tenha alguma dúvida em relação a está pesquisa ou queira dar sua opinião sobre entrar em contato pelo meu linkedin: https://www.linkedin.com/in/diogo-santos-0a8730372/, já caso tenha interesse em obter todos conteúdo envolvendo está pesquisa basta baixar o conteúdo do repositório github: https://github.com/Diogonogueirasantos/Influencia-gerada-por-Redes-sociais.

Entendendo o Dataset

Toda análise de dados tem como parte inicial entender a estrutura de dados que irá ser usada no projeto/pesquisa[1], e por fim "ajustar" os dados para que todos os tipos atribuídos a variáveis estejam adequados às respectivas variáveis. Portanto nesta parte inicial irei fazer alguns ajustes no dataset, seja traduzir variáveis ou simplesmente ajustá-las para o padrão definido no dicionário, que foi previamente concebido antes de sequer pensar em **tratamento de dados** no quesito mais avançado.

Para que serve um dicionário?

De acordo com que lê no livro: **Estatística e Ciência de Dados,** onde no segundo capitulo temos uma parte dedicada ao processo iniciar para a estruturação do projeto, temos que o dicionário será aquilo que servirá como molde, ou melhor o alicerce para à futura construção, logo em um dicionário teremos a definição dos **rótulos** (nomes que serão definidos para substituir nomes que previamente já foram anexados as variáveis) em seguida será definido o tipo de dado atribuído para cada variável se qualitativa: nominal ou ordinal, ou quantitativa: discreta ou contínua, temos será também as limitações definidas para as variáveis ou até mesmo um padrão para uma medida de mensuração. Com isso temos que o dicionário é uma parte extremamente necessária para o processo de **Padronização** e **Normalização** de dados. Para finalizar está breve sessão irei deixar o link para o dicionario criado para este projeto.

1 O termo "Projeto" e "Pesquisa" ainda me parecem muito vago, portanto será normal ao decorrer usar ambas as terminologias

Tratando dados

Bom, para iniciar o processo de tratamento de dados, verifiquei se todos às variáveis estão de acordo com o tipo definido no dicionário para isso usei a função **dtypes** do pandas, com isso temos o resultado abaixo:

```
df.dtypes
```

```
Idade                                                 int64
Genero                                               object
Tempo de Uso em Redes Sociais                       float64
Numero de Plataformas Usadas                          int64
Frenquencia de Posts                                 object
Frequencia de Posts (Ordinal)                       float64
Frequencia de Verificacao por Notificacoes           object
Frequencia de Verificacao por Notificacoes (Ordinal)    float64
Taxa de Reconhecimento de Dependencia               float64
Cyberbullying Experience                              int64
Taxa de Autoestima Reportada                        float64
Qualidade de Sono                                   float64
Taxa de Ansiedade                                   float64
Fadiga Gerada pelo Uso de Redes Sociais               int64
Status de Saude Mental                               object
Status de Saude Mental (Ordinal)                    float64
dtype: object
```

Logo podemos concluir que sim, de fato todas às variáveis estão de acordo com o padrão definido no dicionário (é valido lembrar que o padrão definido no dicionário a via de regra pode vir ter mudanças de acordo com o andamento do projeto). Agora que já verificamos a natureza dos dados vamos tratar dos valores nulos[2], e para isso, usei a função **isnull** de pandas para entender se, a somatória dos casos valeria o esforço da aplicação do tratamento de dados no quesito da inconsistência de registros ou seja valores nulos ou apenas dropá-los do dataset

2 Apesar de geralmente representarmos os valores nulos atribuindo zeros a estes (sim, assim como você já cometi este mesmo erro), valores nulos representam a "ausência" de valor logo o zero em si não representa uma ausência de valor, como já sabemos o zero é usado para representar ou demarcação posicional numérica.

```
df.isnull().mean().round(2) / 2 * 100
```

```
Age                                     0.0
Gender                                  0.0
Daily Social Media Usage(hours)         0.5
Number of Social Media Platforms        0.0
Frequency of Posts                      0.0
Frequency of Checking Notifications     0.5
Self Reported Addiction Score           0.5
Cyberbullying Experience                0.0
Self Esteem Score                       0.5
Sleep Quality                           0.5
Anxiety Score                           0.5
Social Media Fatigue Score              0.0
Mental Health Status                    0.0
dtype: float64
```

Como podemos ver, em nosso dataset de fato há ocorrências de valores nulos, porém, temos a seguinte questão "Vale mesmo apena tratar 5% dos casos deste dataset?", sinceramente acredito que a solução mais lucida para este caso será excluir todos os 5% destes casos uma vez que, passo a entender que não existe representatividade suficiente para entender que estes dados excluídos de fato venham fazer falta.

Observações no tratamento de dados

Gostaria de fazer algumas observações a respeito da limpeza dos dados, como foi mostrado acima todos os registros[3] nulos foram deletados do dataset e depois disso nem outro tipo de tratamento foi feito no quesito limpeza, como por exemplo verificar a duplicidade ou até mesmo digitações com espaços, por um lado podemos enxergar isso como uma limpeza "preguiçosa" porém como este dataset vem do Kaggle, optei por simplesmente não verificar a duplicidade dos dados uma vez que primeiro, não existe um parâmetro como ID ou algo que se possa identificar o usuário na pesquisa, isso se da pelo simples fato de que se por algum motivo eu ou você caro leito tivéssemos acesso a este tipo de informação estaríamos infringindo regras voltadas para a privacidade dos dados. Um outro motivo para a não aplicação de qualquer outro tipo de técnica para a limpeza dos dados, se da pelo fato de que, no geral a maioria dos datasets vindos do Kaggle já possuem este tipo de tratamento de dados, uma vez que o propósito destes datasets está mais voltado para às aplicações de modelos probabilísticos sejam para estimadores para classificação quanto para regressão. Valido lembrar que não sou nem um especialista no assunto, logo isto não pode ser baseada como uma verdade absoluta ou coisa do tipo, veja está minha explicação como algo do tipo " um simples estudante de análise de dados expressando suas descobertas" acredito que seja uma bela maneira de me exaurir de futuras responsabilidades...

3 O termo registro se referi a cada caso situado no dataset, está terminologia vem do Inglês "record"

Análise Descritiva

Indo direto para o ponto que nos interessa, a análise de dados descritiva, você pode estar se perguntando “então você não irá criar modelos probabilísticos?”, e minha resposta seria algo do tipo “não”, porém, isso não significa que eu simplesmente não queira aplicar métodos probabilísticos para uma **análise preditiva**, o que acontece é que sou realista o suficiente para entender que não possuo conhecimentos necessários para falar sobre modelos de machine learning entre outros, prefiro reconhecer está minha falha ao invés de tentar parecer um guru apenas para ganhar algum tipo de reconhecimento. Logo, tudo que veremos a seguir se tratará da análise de dados descritiva.