Projeto - ML RSP DataLab

O Projeto será desenvolvido inicialmente nas seguintes etapas:

- Extração
- Transformação
- Carga (load)
- Desenvolvimento do Machine Learning
- Teste e Implementação do Código
- Implementação da interface gráfica para utilização do usuário final

Esse Arquivo deve ser atualizado a cada etapa do projeto.

Arquitetura / Descrição / Aparência / Formato do DataFrame (DF)

O DataFrame na sua versão final terá aproximadamente um total de 182.500 dados:

- 100 colunas
- 1825 linhas

Os dados serão de 01/01/2017 a 31/12/2019 e 01/01/2021 a 31/12/2023

Primeiramente iremos buscar um método de extração de dados, utilizando o processo ETL.

ETL é uma sigla em inglês que representa as três fases principais do processo de integração de dados em um sistema de gerenciamento de banco de dados ou em um data warehouse: Extract (Extração), Transform (Transformação) e Load (Carga). É uma abordagem comum para coletar, preparar e carregar dados de várias fontes em um formato consolidado e pronto para análises.

Breve explicação de cada fase do processo ETL:

- Extract (Extração): Nesta fase, os dados são coletados de diversas fontes, como bancos de dados, arquivos, APIs ou sistemas externos. Os dados são extraídos dessas fontes e movidos para um ambiente de preparação de dados, geralmente chamado de "área de stage" ou "staging area". A extração pode envolver a leitura de dados brutos ou estruturados, a aplicação de filtros e a validação dos dados para garantir sua integridade.

- Transform (Transformação): Após a extração, os dados são transformados para atender aos requisitos de qualidade e integração do sistema de destino. Essa fase pode incluir várias atividades, como limpeza de dados, enriquecimento de dados, padronização de formatos, agregação, normalização, deduplicação e cálculos de derivados. É nesta fase que os dados são preparados para análises e são transformados em um formato adequado para carregamento no sistema de destino.

- Load (Carga): Depois de extraídos e transformados, os dados são carregados no sistema de destino, que pode ser um banco de dados, um data warehouse ou outra plataforma de armazenamento. Os dados são inseridos na estrutura de armazenamento apropriada, seja em tabelas, cubos ou outros formatos, para que possam ser acessados e analisados pelos usuários finais.

O processo ETL é fundamental para garantir a integridade, qualidade e consistência dos dados em um sistema de gerenciamento de banco de dados ou data warehouse. Ele permite a coleta, preparação e integração de dados de várias fontes em um formato padronizado e pronto para análises, facilitando a tomada de decisões informadas com base nos dados disponíveis.

Exemplo da Arquitetura Final do DF

Coluna 1	Coluna 2	Coluna 3	Coluna 4	Coluna 5	Coluna 6	Coluna 7	Coluna 8	Coluna 9	Coluna 10	...	Coluna 97	Coluna 98	Coluna 99	Coluna 100
Index	Nº tráfegos	Dia/MM/AA	Dia Semana, FDS ou Feriado (FER)?	00:00	00:15	00:30	00:45	01:00	01:15	...	23:00	23:15	23:30	23:45
D001Y01	602	01/01/2017	FER	7	8	5	4	2	3	4	7	2	4	3
D002Y01	784	02/01/2017	SEG	5	2	3	5	5	8	5	7	3	6	5
D003Y01	1001	03/01/2017	TER	4	2	7	5	8	7	3	3	2	8	2
D004Y01	510	04/01/2017	QUA	2	7	7	3	8	8	2	8	4	5	8
D005Y01	856	05/01/2017	QUI	8	7	7	4	7	8	2	8	8	4	2
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
D363Y01	1103	29/12/2017	SEX	7	5	8	7	3	5	5	7	3	8	8

D364Y01	1132	30/12/2017	SÁB	4	3	6	3	2	4	2	4	8	5	4
D365Y01	935	31/12/2017	FER	7	3	2	8	3	8	6	6	7	8	4
D001Y02	710	01/01/2018	FER	7	6	8	2	2	6	8	7	6	6	4
D002Y02	677	02/01/2018	TER	6	7	4	2	3	8	3	7	6	7	5
D003Y02	1170	03/01/2018	QUA	8	4	7	8	8	2	8	8	5	7	5
D004Y02	557	04/01/2018	QUI	5	4	3	3	3	6	7	3	7	4	6
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
D363Y05	1137	29/12/2023	SEX	7	8	4	8	6	3	8	5	7	5	4
D364Y05	873	30/12/2023	SÁB	3	6	2	3	6	5	6	7	8	3	7
D365Y05	755	31/12/2023	FER	8	7	8	2	2	5	8	5	7	8	7

Com esses dados iremos iniciar o processo de Transformação, onde analisaremos o DF e faremos ajustes de linhas e colunas além de transformação dos dados em códigos int8 int64 , string, etc.

Por fim, faremos a carga (Load) do DF no google colab para utilizarmos ferramentas de Machine Learning para leitura e aprendizado de máquina.

Após sucesso de no mínimo 90% de rating, que é um sucesso razoável, pois o dia tem 24 horas - de 15 em 15 minutos temos 96 colunas, ou seja, um rating de 90% nos dá 86 colunas de previsibilidade correta, sendo que as outras 10 colunas provavelmente a IA errará um peso para cima ou para baixo. Isso dá uma ótima predição a ser utilizada pelo supervisor no dia-a-dia para a distribuição da carga horária da equipe.

Esse Arquivo deve ser atualizado a cada etapa do projeto:

- Extração
- Transformação
- Carga (load)
- Desenvolvimento do Machine Learning
- Teste e Implementação do Código
- Implementação da interface gráfica para utilização do usuário final