SERVIÇO PÚBLICO FEDERAL · MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE VIÇOSA · UFV
CAMPUS FLORESTAL

Projeto - Parte 4

Ciência de dados

BRUNO MARRA DE MELO [3029]
VINÍCIUS GABRIEL J. ALMEIDA [3023]

Florestal, MG
2020

Sumário

1. Fase 1

O primeiro passo para realização do trabalho, foi a definição do tema que o grupo tinha interesse em abordar. No caso, após uma discussão, foi definido que seria sobre a distribuição de renda no Brasil. Para tal, uma lista de dados do IBGE foi levantada. Ao final, apenas alguns deles foram utilizados para o trabalho.

Feito isso, o primeiro passo foi o levantamento das perguntas. Foram levantadas 20 perguntas, nas quais cerca de 10 delas foram respondidas ao longo das demais fases do trabalho e serão abordadas em suas respectivas fases.

Após o levantamento das perguntas que desejávamos responder, foi feito o levantamento dos dados que poderiam ser interessantes para extrair e responder às perguntas propostas. Os *datasets* que foram utilizados então foram do IBGE, referentes aos consumos alimentares das famílias baseados em sua renda por estado brasileiro. Um outro *dataset* de interesse do grupo foi relativo aos consumos dos brasileiros no geral, também relativos a renda da população. Um exemplo dos dados em questão é exemplificado pela Figura 1.

Tabela 3.1 - Aquisição alimentar domiciliar *per capita* anual, por Unidades da Federação, segundo os produtos - Região Norte - período 2017-2018

(continua)

Produtos	Aquisição alimentar domiciliar *per capita* anual (kg)							
	Região Norte	Unidades da Federação						
		Rondonia	Acre	Amazonas	Roraima	Pará	Amapá	Tocantins
Cereais e leguminosas	26,644	30,856	29,392	17,211	37,779	27,170	27,908	37,050
Cereais	21,324	25,521	24,761	12,850	29,206	21,577	19,950	32,910
Arroz não especificado	1,971	1,096	3,138	0,962	1,222	2,954	1,092	0,187
Arroz polido	17,942	23,069	18,952	10,849	26,882	16,930	18,162	32,064
Milho em grão	0,840	0,744	1,244	0,698	0,265	1,108	0,256	0,112
Milho verde em conserva	0,135	0,263	0,131	0,031	0,135	0,165	0,067	0,130
Milho verde em espiga	0,180	0,078	0,636	0,021	0,004	0,224	0,108	0,299
Outros	0,257	0,271	0,658	0,290	0,697	0,196	0,264	0,117
Leguminosas	5,320	5,335	4,631	4,361	8,572	5,593	7,958	4,140
Feijão-fradinho	0,160	0,038	0,234	0,063	0,106	0,237	0,094	0,131
Feijão-jalo	0,211	-	-	0,433	-	0,205	0,040	0,204
Feijão-manteiga	0,096	0,028	0,080	0,163	-	0,086	0,131	0,075
Feijão-mulatinho	0,482	1,094	0,347	0,489	0,640	0,368	0,726	0,292
Feijão-preto	0,806	0,377	0,126	0,522	0,135	1,220	1,225	0,109
Feijão-rajado	2,282	3,002	1,965	1,590	3,546	2,265	4,470	1,916
Feijão-roxo	0,005	-	-	0,009	-	-	0,018	0,028
Outros feijões	0,938	0,255	1,486	0,859	2,926	0,960	1,034	0,783
Outras	0,339	0,540	0,393	0,234	1,220	0,252	0,220	0,602
Hortaliças	11,594	16,979	14,608	9,809	14,046	10,692	17,044	9,612
Hortaliças folhosas e florais	1,367	2,314	2,486	0,886	1,246	1,163	3,704	0,819
Acelga	0,017	0,078	0,035	0,005	0,020	0,005	0,033	0,016
Agrião	0,003	0,013	-	0,005	-	-	0,006	-
Alface	0,197	0,489	0,362	0,130	0,220	0,145	0,240	0,204
Cheiro-verde	0,351	0,176	0,529	0,286	0,475	0,356	1,173	0,118
Couve	0,212	0,341	0,676	0,149	0,233	0,138	0,825	0,038
Couve-brócolis	0,046	0,081	0,045	0,036	0,004	0,052	0,054	0,010
Couve-flor	0,017	0,101	0,049	0,005	-	-	0,049	0,016
Repolho	0,376	0,822	0,475	0,202	0,267	0,329	0,655	0,408

Figura 1 - Exemplo de dados utilizados

2. Fase 2

Os dados foram retirados do site do IBGE através da referência [1], esses estão armazenados em planilhas excel, que contém informações para cada estado do Brasil. Essas planilhas se encontram no diretório: dados-ibge/tabelas_unidades_da_federacao_xls_20191108 .

Cada estado possui uma tabela com 6 abas, das quais apenas a primeira e a terceira foram relevantes para nosso trabalho. Sendo assim, para cada estado foi feita uma conversão dessas duas abas para CSV, pois neste formato é melhor para extrair as informações. O resultado deste pré-processamento se encontra em: dados-limpos/aquisicao_por_classe_rendimento_e_estado/SIGLA_ESTADO/dados-ibge/ .

Com isso, com esses dados em CSV foi necessário apenas organizar os *headers* e copiar as linhas de interesse para um novo arquivo, feito à mão, que se encontra em: dados-limpos/aquisicao_por_classe_rendimento_e_estado/SIGLA_ESTADO/dados-limpos/ . Essa escolha foi feita para que pudéssemos partir de um *dataset* já formatado e sem a necessidade de fazer a limpeza via código, visto que foi trivial fazer à mão.

Com isso, basta utilizar este *dataset* como *dataframe* para então realizar as análises seguintes.

3. Fase 3

Já na terceira fase, utilizamos os dados já limpos da segunda fase para realizar uma análise exploratória dos dados, bem como já tentar responder algumas das perguntas realizadas ainda na Fase 1.

Sendo assim, primeiro para o *dataset* de despesas com alimentação, foi feita a verificação do top 5 de alimentos consumidos de cada classe social em Minas Gerais, e, com isso, foi possível começar a responder algumas perguntas ainda nessa fase. Por exemplo, pode-se verificar claramente o crescimento do item alimentação fora do domicílio à medida que a renda aumenta, por exemplo. A Figura 2 representa esse detalhe.

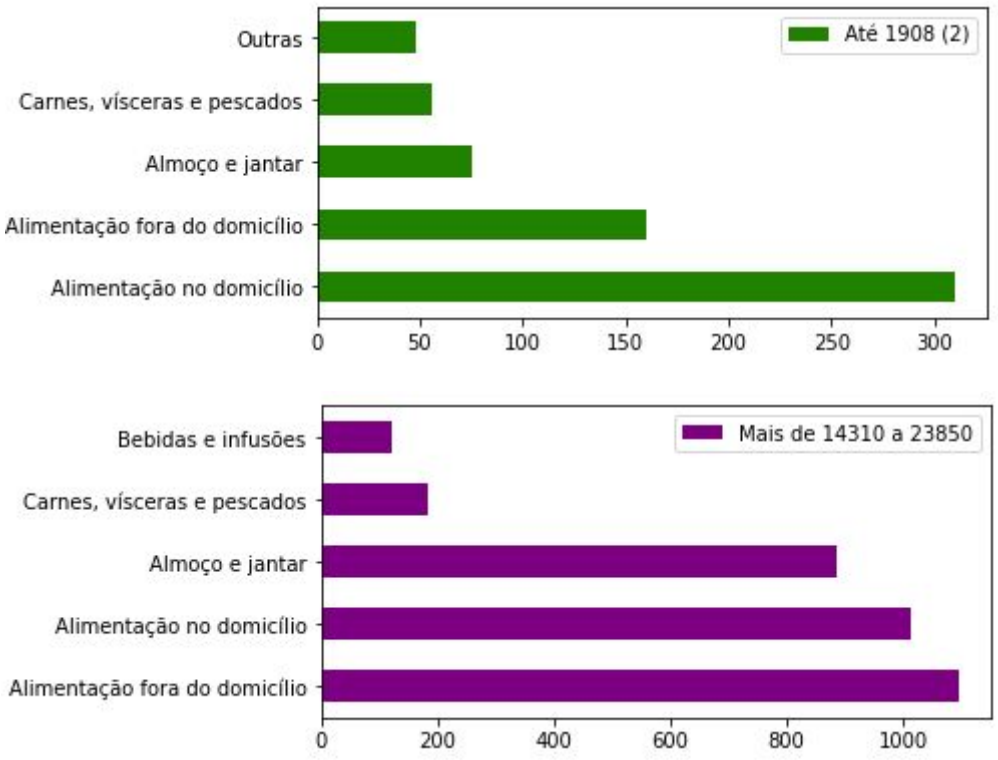

Figura 2 - Diferença de alimentação de acordo com a renda

Além disso, outros padrões importantes foram identificados nessa fase, como por exemplo ser possível aplicar publicidade direcionada para populações, sabendo a renda da população em uma região, bem como quais classes de produtos ditam tendência no mercado alimentício em termos de quantidade de consumo. Podemos perceber também que o consumo de aves e ovos no estado do PA é muito maior que nos demais estados, bem como possui maior valor consumido em suas alimentações em casa.

O outro *dataset* utilizado, com os mesmos padrões do anterior, porém com dados diferentes foi o de despesas gerais, onde se encontram quais os gastos mais comuns da população baseadas em sua renda. Analogamente, foi levantado o top 5 de cada uma das rendas em Minas Gerais, e assim como o anterior, foi possível encontrar alguns padrões de consumo úteis, como por exemplo a aparição do item aluguel como grande gasto nas classes mais baixas e nem aparecendo na classe mais favorecida financeiramente, como pode ser observado pela Figura 3.

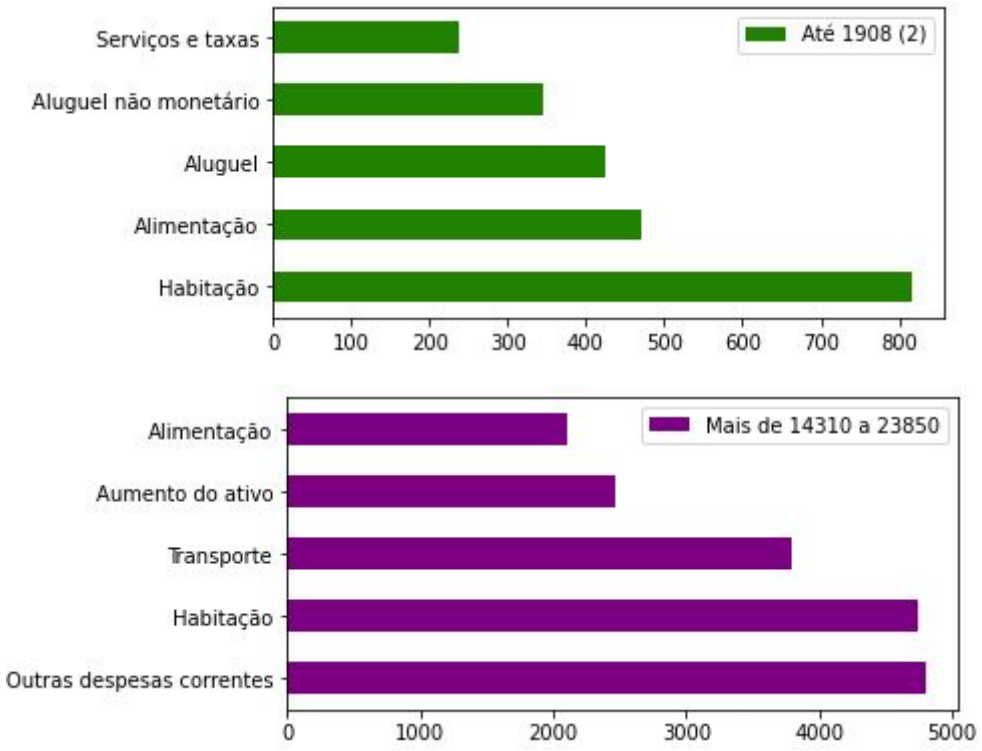

Figura 3 - Diferença de consumo de acordo com a renda

Da mesma forma que a análise feita ao primeiro conjunto de dados, também é possível realizar e aplicar publicidade exclusiva e direcionada para produtos em regiões específicas, somente se baseando nos dados de consumo do senso, bem como sabendo direcionar para o público-alvo correto.

Além disso, é possível analisar fortes tendências de mercado de acordo com a região que é analisada, por exemplo, o consumo por meios de transporte no estado do Rio Grande do Sul é bem maior que no estado do Pará. Dessa forma, pode-se notar uma tendência do mercado automobilístico no estado. Podemos notar também, por exemplo, que o investimento em ativos para rendas maiores no RS é mais evidente que em MG e SP. Analogamente, os aluguéis em SP são muito mais comuns para classes mais altas do que no RS e em MG.

4. Fase 4

Para a realização da quarta e última etapa foi proposta uma última análise dos dados, neste caso, a predição. Com a base de dados já preparada e com melhorias incrementadas gradualmente com as outras entregas, nesta etapa partimos de um *dataframe* já formatado e limpo.

Antes de aplicarmos as possíveis técnicas estudadas na disciplina discutimos sobre o que seria coerente e viável de se utilizar com nosso *dataset*. Dentre as opções, vimos que muitas não seriam viáveis. Começando pelo **aprendizado não supervisionado**, tem-se o seguinte: seria muito útil para agrupar informações não relacionadas, a princípio, em possíveis grupos promissores. Entretanto, nossos dados já vieram "geolocalizados" de forma aproximada, ou seja, por estados brasileiros. Logo, essa poderia ser uma das classificações relevantes. Outras opções para este tipo de aprendizado poderia estar relacionada aos tipos de famílias com consumos e localização. Porém, nossa base de dados não possui dados com localização precisa e também não divide as famílias dentro de cada classe social. Dessa forma, não é possível saber quais os produtos exatos que cada família consome. Isso se torna um grande obstáculo, pois não é possível conhecer cada família individualmente e tentar agrupá-las de forma coerente. Sendo assim, descartamos esta opção de aprendizado não supervisionado.

Outra opção de análise são as **regras de associação**. O seu uso para esse *dataset* seria muito relevante e interessante de analisar. Entretanto, nos deparamos com os mesmos problemas: os dados das famílias não estão separados individualmente ou então por subclasses que consomem os mesmos tipos de produtos. Com isso, não faz sentido aplicar essa regra, pois todas as linhas de classes sociais e estados possuem valores misturados e preenchendo praticamente todas as colunas. Logo, geraria associações para inúmeros produtos (se não todos) juntos, o que não representa a realidade.

Uma outra possibilidade seria o uso de **regressão linear** para tentar descrever parte do consumo de cada família. Entretanto, nos deparamos com os mesmos problemas mencionados acima, as famílias não possuem dados individuais. Desta vez, mesmo sem este tipo de informação ainda é possível tentar fazer uma análise mais detalhada (aproximada), já que não possuímos muitos detalhes sobre as famílias individualmente. Para isso, foi feita uma análise de regressão linear para cada classe social dos dois *datasets* (Geral e Alimentício). Como todos os dados que explicam o valor total estão no *dataset* é de se esperar que o R2 (score) seja 1 (100% explicado). Apenas para demonstração, será apresentada a equação formada para um das seis classes sociais. A classe social escolhida é: "Mais de 1908 a 2862" e o *dataset* é o de gastos gerais, Figura 4.

```
R2 (score): 1.0

Intercept: 360.47547

Equação do 'Total gasto por classes' em termos de todos os atributos analisados:

  f(x) = 360.47547 + (1.27970 * 'Alimentação') + (-0.41236 * 'Habitação') + (2.70866 * 'Aluguel') + (1.37569 * 'Aluguel monetário') + (0.83935 * 'Aluguel não monetário') + (0.43709 * 'Condomínio') + (2.63586 * 'Serviços e taxas') + (-0.23281 * 'Energia elétrica') + (-0.09799 * 'Telefone fixo') + (0.28810 * 'Telefone Celular') + (0.56844 * 'Pacote de telefone, TV e Internet') + (-0.62182 * 'Gás doméstico') + (-0.37654 * 'Água e esgoto') + (0.37408 * 'Outros') + (1.94951 * 'Manutenção do lar') + (0.28437 * 'Artigos de limpeza') + (1.28867 * 'Mobiliários e artigos do lar') + (0.93835 * 'Eletrodomésticos') + (0.06793 * 'Consertos artigos do lar') + (1.73781 * 'Vestuário') + (0.23325 * 'Roupa de homem') + (1.00231 * 'Roupa de mulher') + (0.37892 * 'Roupa de criança') + (0.35859 * 'Calçados e apetrechos') + (0.10641 * 'Jóias e bijuterias') + (0.13290 * 'Tecidos e armarinhos') + (1.08802 * 'Transporte') + (0.26389 * 'Urbano') + (0.67790 * 'Gasolina - veículo próprio') + (0.72658 * 'Álcool - veículo próprio') + (0.58915 * 'Manutenção e acessórios') + (0.91886 * 'Aquisição de veículos') + (0.28668 * 'Viagens esporádicas') + (0.11503 * 'Outras') + (-0.16393 * 'Higiene e Cuidados Pessoais') + (-0.15789 * 'Perfume') + (0.16362 * 'Produtos para cabelo') + (0.19730 * 'Sabonete') + (0.23903 * 'Instrumentos e produtos de uso pessoal') + (0.00323 * 'Assistência a saúde') + (0.84283 * 'Remédios') + (0.07449 * 'Plano/Seguro saúde') + (0.14834 * 'Consulta e tratamento dentário') + (-0.08539 * 'Consulta médica') + (1.02357 * 'Tratamento médico e ambulatorial') + (2.19022 * 'Serviços de cirurgia') + (0.58889 * 'Hospitalização') + (-0.82208 * 'Exames diversos') + (0.31845 * 'Material de tratamento') + (-0.16501 * 'Outras') + (-0.05292 * 'Educação') + (0.29913 * 'Cursos regulares') + (-0.05185 * 'Cursos superiores') + (-0.01321 * 'Outros cursos e atividades') + (0.07949 * 'Livros didáticos e revistas técnicas') + (0.11503 * 'Artigos escolares') + (-0.16393 * 'Outras') + (-0.15789 * 'Recreação e cultura') + (0.16362 * 'Brinquedos e jogos') + (0.19730 * 'Celular e acessórios') + (0.23903 * 'Periódicos, livros e revistas não didáticos') + (0.00323 * 'Recreações e esportes') + (1.16273 * 'Outras') + (0.54097 * 'Fumo') + (1.15761 * 'Serviços pessoais') + (0.21307 * 'Cabeleireiro') + (-0.25963 * 'Manicuro e pedicuro') + (0.17438 * 'Consertos de artigos pessoais') + (0.11503 * 'Outras') + (-0.16393 * 'Despesas diversas') + (-0.15789 * 'Jogos e apostas') + (0.16362 * 'Comunicação') + (0.19730 * 'Cerimônias e festas') + (0.23903 * 'Serviços profissionais') + (0.00323 * 'Imóveis de uso ocasional') + (1.65810 * 'Outras') + (0.11233 * 'Outras despesas correntes') + (1.45001 * 'Impostos') + (0.06512 * 'Contribuições trabalhistas') + (0.36977 * 'Serviços bancários') + (0.11503 * 'Pensões, mesadas e doações') + (-0.16393 * 'Previdência privada') + (-0.15789 * 'Outras') + (0.16362 * 'Aumento do ativo') + (0.19730 * 'Imóvel (aquisição)') + (0.23903 * 'Imóvel (reforma)') + (0.00323 * 'Outros investimentos') + (0.26874 * 'Diminuição do passivo') + (0.25728 * 'Empréstimo') + (-0.00270 * 'Prestação de imóvel') + (-0.02160 * 'Tamanho médio da família')

Coeficiente de determinação é: 1.00000, ou seja, 0.00000% do 'Total gasto por classes' não é explicada pelos atributos utilizados
```

Figura 4 - Equação para classe social "Mais de 1908 a 2862" e *dataset* "Geral".

Por se tratar de uma equação que representa valores contínuos, o significado de cada termo multiplicado é: um peso para cada categoria vezes o valor gasto nessa categoria. Acredita-se que estes valores de pesos não ficaram iguais a 1 porque foi analisado o consumo para todos os 27 estados brasileiros juntos para cada classe social. Com isso, cada estado possui gastos diferentes e totais diferentes. Sendo assim, ao nosso ver, essa equação também não é relevante e significativa. Pois, como depende dos gastos individuais de cada família, ao se ter isso não é preciso usar uma equação para achar o total consumido, pois basta apenas somar estes valores. Com isso, acreditamos que foi uma tentativa válida para o que tínhamos de dados, mas que não será útil.

Por fim, tem-se o **aprendizado supervisionado**, que para o nosso *dataset* foi o mais relevante e significativo.

Partindo do *dataframe* construído na etapa 3, temos as seguintes informações, Figura 5.

	Tipos de despesa, número e tamanho médio das famílias	Até 1908 (2)	Mais de 1908 a 2862	Mais de 2862 a 5724	Mais de 5724 a 9540	Mais de 9540 a 14310	Mais de 14310 a 23850
0	Alimentação	470.26	682.91	965.76	1183.04	1511.11	2111.32
1	Habitação	817.04	1148.45	1658.10	2278.54	3102.18	4744.36
2	Aluguel	423.40	581.47	799.82	1098.35	1430.00	2098.02
...	...	...	...	...	...	...	...
90	Número de famílias	1478894.00	2470800.00	1092145.00	384637.00	207088.00	123662.00
91	Tamanho médio da família	2.69	3.13	3.48	3.11	2.93	3.02
92	Total gasto por classes	4723.72	7618.75	13440.48	19663.04	29365.22	48634.53

93 rows × 7 columns

Figura 5 - *Dataframe* da etapa 3.

Contudo, para um algoritmo supervisionado, este *dataframe* não possui informações relevantes e significativas para serem utilizadas como rótulos. Sendo assim, foi necessário fazer algumas modificações no seu formato para possibilitar o uso de rótulos. Com base nisso, utilizamos a ideia de classificar dados para classes sociais, assim como questionado na primeira etapa do trabalho.

Antes de fazer este ajuste, foi necessário apenas fazer uma conversão na representação dos dados. Por variar muito dentro do total consumido por cada classe/estado, achamos mais relevante usar informações percentuais, com isso seria possível normalizar os dados e criar proporções mais representativas, Figura X. Em seguida, para implementar a ideia de rótulos foi utilizado um comando muito útil do *framework* pandas, o **df.transpose()**. Com isso, ele pega o *dataframe* atual e inverte suas linhas com colunas, gerando o seguinte resultado, Figuras 6 e 7.

	Tipos de despesa, número e tamanho médio das famílias	Até 1908 (2)	Mais de 1908 a 2862	Mais de 2862 a 5724	Mais de 5724 a 9540	Mais de 9540 a 14310	Mais de 14310 a 23850
0	Alimentação	0.10	0.09	0.07	0.06	0.05	0.04
1	Habitação	0.17	0.15	0.12	0.12	0.11	0.10
2	Aluguel	0.09	0.08	0.06	0.06	0.05	0.04
...	...	...	...	...	...	...	...
90	Número de famílias	1478894.00	2470800.00	1092145.00	384637.00	207088.00	123662.00
91	Tamanho médio da família	2.69	3.13	3.48	3.11	2.93	3.02
92	Total gasto por classes	1	1	1	1	1	1

93 rows × 7 columns

Figura 6 - *Dataframe* da etapa 3.

	Alimentação	Habitação	Aluguel	Aluguel monetário	Aluguel não monetário	Condomínio	Serviços e taxas	Energia elétrica	Telefone fixo	Telefone Celular	...	Aumento do ativo	Imóvel (aquisição)	Imóvel (reforma)	Outros investimentos	Diminuição do passiv
0	0.10	0.17	0.09	0.02	0.07	0.00	0.05	0.02	0.00	0.01	...	0.01	0.00	0.01	0.00	0.0
1	0.09	0.15	0.08	0.01	0.07	0.00	0.04	0.01	0.00	0.01	...	0.01	0.00	0.01	0.00	0.0
2	0.07	0.12	0.06	0.01	0.05	0.00	0.04	0.01	0.00	0.01	...	0.02	0.01	0.01	0.00	0.0
3	0.06	0.12	0.06	0.01	0.05	0.00	0.03	0.01	0.00	0.01	...	0.02	0.01	0.01	0.00	0.0
4	0.05	0.11	0.05	0.01	0.04	0.01	0.02	0.01	0.00	0.00	...	0.02	0.01	0.01	0.00	0.0
5	0.04	0.10	0.04	0.00	0.04	0.01	0.02	0.00	0.00	0.00	...	0.05	0.04	0.01	0.00	0.0

6 rows × 93 columns

Figura 7.1 - *Dataframe* invertido (Parte 1).

luguel não etário	Condomínio	Serviços e taxas	Energia elétrica	Telefone fixo	Telefone Celular	...	Aumento do ativo	Imóvel (aquisição)	Imóvel (reforma)	Outros investimentos	Diminuição do passivo	Empréstimo	Prestação de imóvel	Tamanho médio da família	Total gasto por classes	Classe social
0.07	0.00	0.05	0.02	0.00	0.01	...	0.01	0.00	0.01	0.00	0.01	0.01	0.00	2.69	1	Até 1908 (2)
0.07	0.00	0.04	0.01	0.00	0.01	...	0.01	0.00	0.01	0.00	0.01	0.01	0.00	3.13	1	Mais de 1908 a 2862
0.05	0.00	0.04	0.01	0.00	0.01	...	0.02	0.01	0.01	0.00	0.02	0.01	0.01	3.48	1	Mais de 2862 a 5724
0.05	0.00	0.03	0.01	0.00	0.01	...	0.02	0.01	0.01	0.00	0.02	0.01	0.00	3.11	1	Mais de 5724 a 9540
0.04	0.01	0.02	0.01	0.00	0.00	...	0.02	0.01	0.01	0.00	0.02	0.01	0.01	2.93	1	Mais de 9540 a 14310
0.04	0.01	0.02	0.00	0.00	0.00	...	0.05	0.04	0.01	0.00	0.01	0.00	0.01	3.02	1	Mais de 14310 a 23850

Figura 7.2 - *Dataframe* invertido (Parte 2).

Dessa forma, podemos observar que os dados já estão representados por suas respectivas colunas de consumos percentuais, e a última coluna “Classe social” com o rótulo que estávamos procurando.

Como pode ser observado, esse *dataframe* é apenas para um estado, com uma linha para cada classe social. O *dataframe* que utilizamos para a predição é uma pequena modificação que inclui todos os estados brasileiros, coluna “Estado”, e com mais linhas, 6 linhas para cada um dos 27 estados, resultando em 162 linhas, Figura 8 e 9.

	Alimentação	Habitação	Aluguel	Aluguel monetário	Aluguel não monetário	Condomínio	Serviços e taxas	Energia elétrica	Telefone fixo	Telefone Celular	...	Imóvel (aquisição)	Imóvel (reforma)	Outros investimentos	Diminuição do passivo	Empr
0	0.12	0.15	0.08	0.00	0.08	0.00	0.04	0.02	0.00	0.01	...	0.00	0.01	0.00	0.01	
1	0.10	0.14	0.07	0.00	0.07	0.00	0.04	0.02	0.00	0.01	...	0.00	0.01	0.00	0.03	
2	0.06	0.12	0.06	0.00	0.06	0.00	0.03	0.01	0.00	0.01	...	0.02	0.02	0.00	0.04	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
159	0.05	0.12	0.06	0.01	0.05	0.00	0.04	0.02	0.00	0.00	...	0.00	0.00	0.00	0.02	
160	0.04	0.11	0.04	0.00	0.04	0.00	0.03	0.01	0.00	0.00	...	0.00	0.00	0.00	0.01	
161	0.03	0.11	0.04	0.01	0.03	0.00	0.02	0.01	0.00	0.00	...	0.03	0.02	0.00	0.06	

162 rows × 94 columns

Figura 8 - *Dataframe* invertido para todos os estados (Parte 1).

Aluguel não netário	Condomínio	Serviços e taxas	Energia elétrica	Telefone fixo	Telefone Celular	...	Imóvel (aquisição)	Imóvel (reforma)	Outros investimentos	Diminuição do passivo	Empréstimo	Prestação de imóvel	Tamanho médio da família	Total gasto por classes	Estado	Classe social
0.08	0.00	0.04	0.02	0.00	0.01	...	0.00	0.01	0.00	0.01	0.01	0.00	3.72	1	AC	Até 1908 (2)
0.07	0.00	0.04	0.02	0.00	0.01	...	0.00	0.01	0.00	0.03	0.03	0.00	4.07	1	AC	Mais de 1908 a 2862
0.06	0.00	0.03	0.01	0.00	0.01	...	0.02	0.02	0.00	0.04	0.04	0.00	3.99	1	AC	Mais de 2862 a 5724
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.05	0.00	0.04	0.02	0.00	0.00	...	0.00	0.00	0.00	0.02	0.02	0.01	3.48	1	TO	Mais de 5724 a 9540
0.04	0.00	0.03	0.01	0.00	0.00	...	0.00	0.00	0.00	0.01	0.01	0.00	3.43	1	TO	Mais de 9540 a 14310
0.03	0.00	0.02	0.01	0.00	0.00	...	0.03	0.02	0.00	0.06	0.04	0.02	3.01	1	TO	Mais de 14310 a 23850

Figura 9 - *Dataframe* invertido para todos os estados (Parte 2).

Logo, com este *dataframe* (consumo geral) preparado, basta utilizá-lo como base de dados para algoritmos preditivos. Neste trabalho avaliamos 5 algoritmos: KNN, Naive Bayes, Árvore de Decisão, SVM e MLP. A divisão entre testes e treinos foi de 20% para testes e 80% para treino. Dentre estes algoritmos avaliados, o que obteve melhor resultado foi o algoritmo de Árvore de Decisão, com os seguintes resultados: precisão média de 71% e revocação média de 70%, Figura 10.

Árvore de Decisão

	precision	recall	f1-score	support
Até 1908 (2)	0.80	1.00	0.89	4
Mais de 14310 a 23850	0.50	0.67	0.57	3
Mais de 1908 a 2862	1.00	0.67	0.80	3
Mais de 2862 a 5724	0.73	0.89	0.80	9
Mais de 5724 a 9540	0.83	0.50	0.62	10
Mais de 9540 a 14310	0.40	0.50	0.44	4
accuracy			0.70	33
macro avg	0.71	0.70	0.69	33
weighted avg	0.73	0.70	0.69	33

Figura 10 - Resultado para algoritmo de Árvore de Decisão (*Dataframe* consumo geral).

Sua matriz de confusão obteve o seguinte resultado, Figura 11.

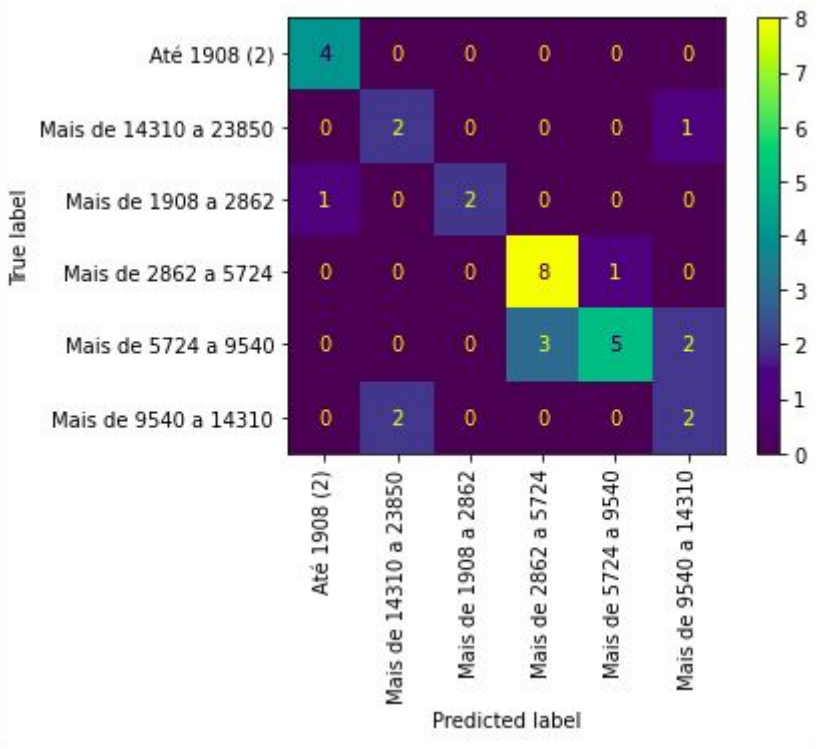

Figura 11 - Matriz de confusão da Árvore de Decisão (*Dataframe* consumo geral).

Para o outro *dataframe* (consumo alimentício) os resultados não foram tão assertivos quanto os de consumo geral. Vale a pena ressaltar que foi aplicado o mesmo critério de conversão dos dados, sendo necessário apenas trocar o diretório para o arquivo CSV com os dados pré-processados para obter o *dataframe* exibido anteriormente. Com isso, após submeter esta base de dados nova aos mesmos algoritmos, os resultados foram inferiores que os obtidos na etapa anterior. O algoritmo que se sobressaiu neste caso foi o de Naive Bayes, com precisão média de 46% e revocação média de 39%, Figura 12.

Naive Bayes

	precision	recall	f1-score	support
Até 1908 (2)	0.30	0.75	0.43	4
Mais de 14310 a 23850	1.00	0.33	0.50	3
Mais de 1908 a 2862	0.13	0.67	0.22	3
Mais de 2862 a 5724	0.00	0.00	0.00	9
Mais de 5724 a 9540	1.00	0.10	0.18	10
Mais de 9540 a 14310	0.33	0.50	0.40	4
accuracy			0.27	33
macro avg	0.46	0.39	0.29	33
weighted avg	0.48	0.27	0.22	33

Figura 12 - Resultado para algoritmo de Naive Bayes (*Dataframe* consumo alimentício).

Sua matriz de confusão obteve o seguinte resultado, Figura 13.

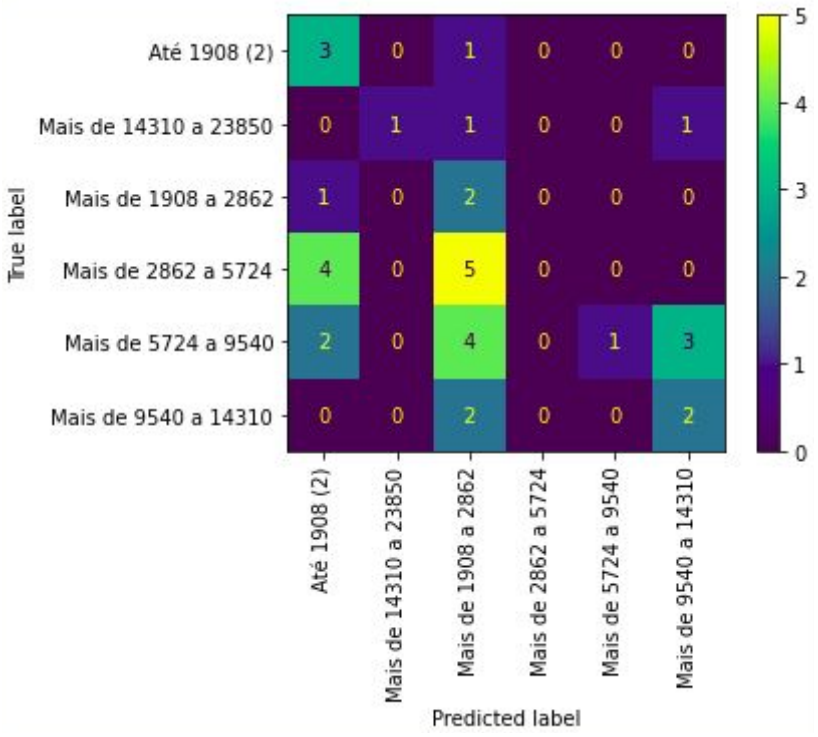

Figura 13 - Matriz de confusão do Naive Bayes (*Dataframe* consumo alimentício).

Com isso, podemos concluir que a análise preditiva foi o melhor resultado obtido para os dados estudados na Fase 4. Mostrando que é possível aplicá-la para tentar reconhecer de qual classe social uma pessoa qualquer é.

Queríamos ter conseguido uma base de dados mais detalhada para poder realizar as outras análises citadas anteriormente, porém, não foi possível. Entretanto, apesar das dificuldades encontradas pelos *datasets* disponíveis, serviu de aprendizado para entender que faz parte não ter todas as informações que precisamos, além de entender a complexidade para consegui-las. Sendo assim, concluímos que dentro do possível grande parte das questões levantadas foram respondidas, mostrando que é possível agregar significados relevantes e positivos para os dados estudados.

5. Referências

[1] Dados IBGE https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html