Kennedy Anderson Guimarães de Araújo

- 24 anos;
- Bacharel em Matemática Industrial, Janeiro/2017;
- Mestrando em Modelagem e Métodos Quantitativos, Março/2019;
- Características e interesses acadêmicos:
  - Aprendizado de Máquina;
  - Mineração de Dados;
  - Otimização Combinatória;
  - Pesquisa Operacional;
  - Modelagem Estatística;
  - Planejamento e Controle de Produção.

Sumário

- Ferramentas utilizadas para o trabalho.
- Pré-processamento dos dados.
- Estatísticas gerais.
- Seleção de características mais importante e análise descritiva.
- Seleção de 1000 clientes para ofertar fatura por e-mail.
- Seleção de 1000 clientes para ofertar SMS alerta.
- Escolha e ajuste de modelos de aprendizado de máquina para prever qual produto ofertar para um certo cliente.

Ferramentas utilizadas para o trabalho.

- Linguagem de programação *Python 3.7* com o IDE *Spyder*.
- Bibliotecas *pandas*, *numpy* e *datetime* para tratamento dos dados.
- Bibliotecas *matplotlib*, *seaborn* e *pylab* para confecção dos gráficos.
- Bibliotecas *sklearn* para ajuste e seleção de modelos de aprendizado de máquina.
- Excel visualização de tabelas e criação de tabelas dinâmicas.

Pré-processamento do dados

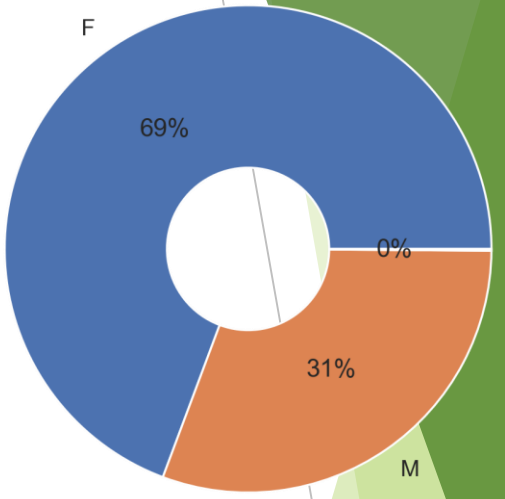
- Transformação de **DataNascimento** coluna **Idade;**
- Tratamento de valores faltantes: substituídos por ‘NaN’;
- Padronizar valores categóricos, caso necessário:
  - Exemplo: verificar se existiam casos como **ACESSÓRIOS/CALÇADOS/ARTIGOS ESPORTIVOS** e **ACESSÓRIOS / CALÇADOS / ARTIGOS ESPORTIVOS**, os dois significariam a mesma coisa, mas teriam valores diferentes nos dados.
  - Não foram constatados casos como este.

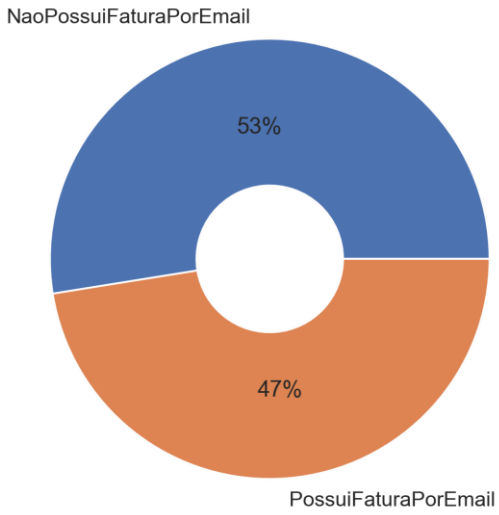
Pré-processamento do dados

- Anomalias nos dados:
  - Clientes com fatura por e-mail sem e-mail cadastrado;
  - Clientes com idade muito avançada ou muito jovens:
    - Exemplo: havia clientes com 2 anos e outros com 120 anos.
  - Remoção de clientes com idade menor que 18 anos e maior que 100 anos;
  - Remoção de clientes com fatura por e-mail, porém sem e-mail cadastrado
  - Detecção de clientes com mesmos dados cadastrais (duplicados).
  - Redução de 67454 para 66289.

Estatísticas gerais

- Clientes: 66289;
- Cidades: 1086;
- Estados: 21;
- Setores: 12;
- Idades: média de 41.54 anos e desvio padrão de 13.31 anos.
- Compras no último ano: média de 7.86 e desvio padrão de 8.53.
- Tempo médio na FortBrasil: média de 6.41 meses e desvio padrão de 2.62 meses.





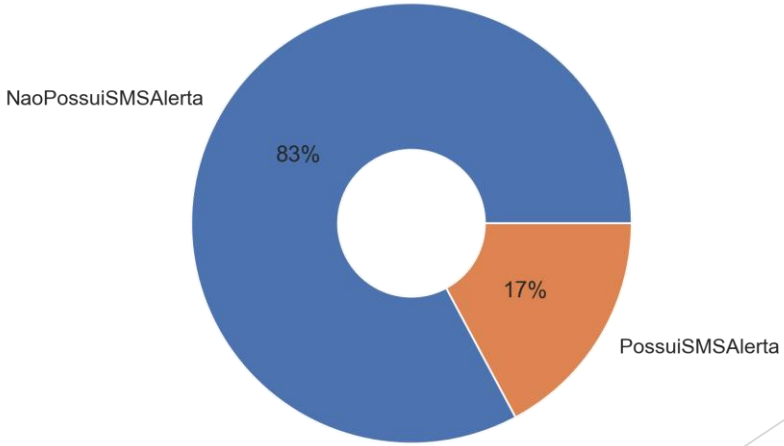
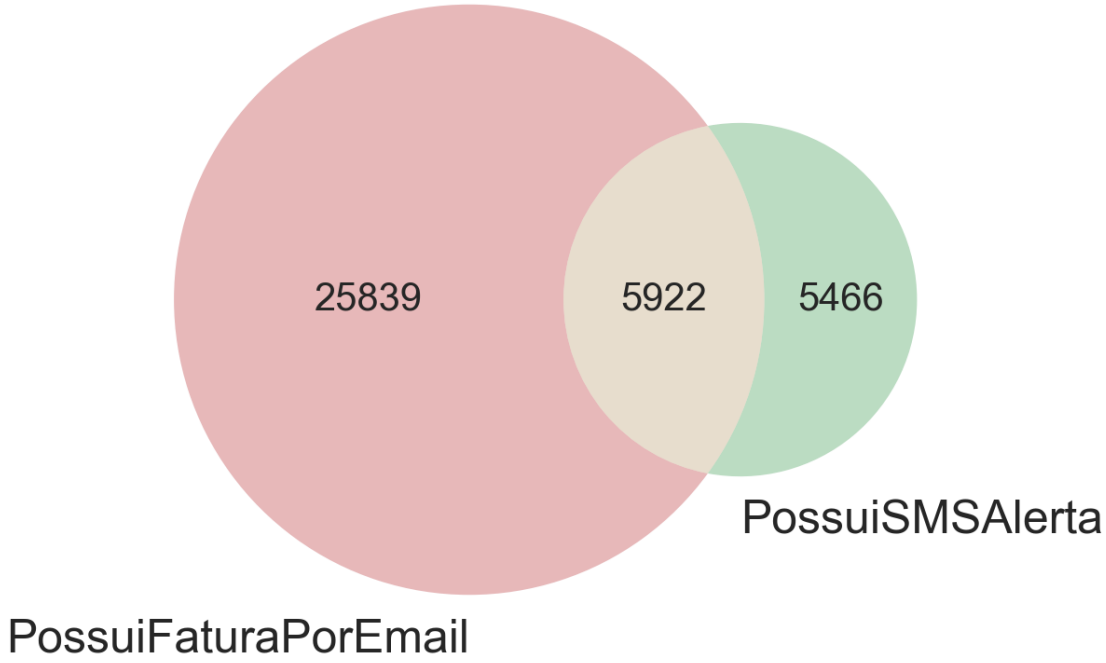


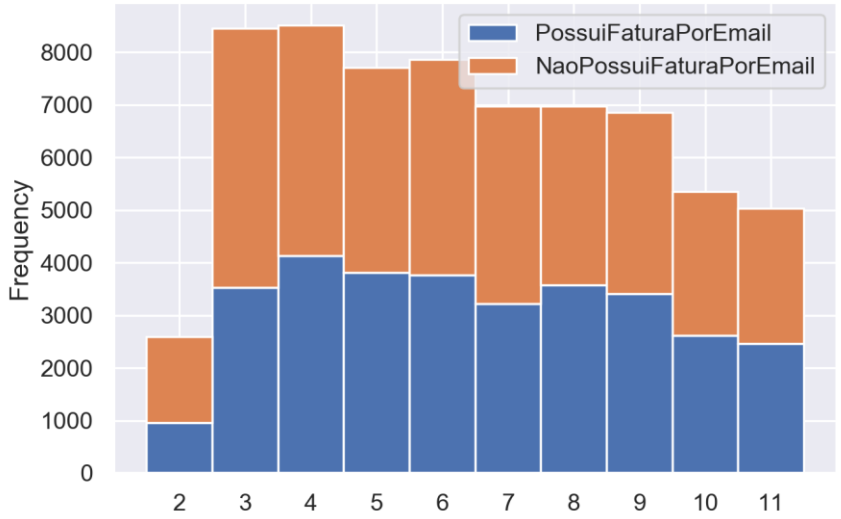
Diagrama de Venn para clientes com fatura por e-mail e SMS alerta

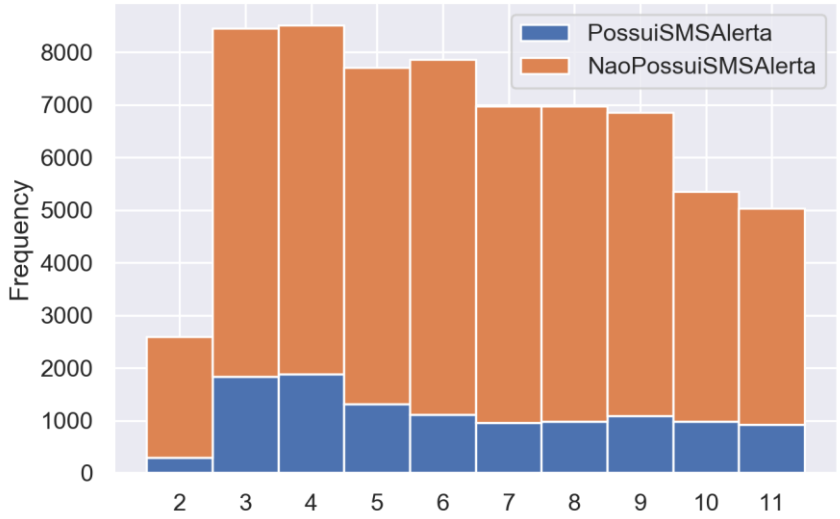


Seleção das características mais importantes para redução da dimensão dos dados

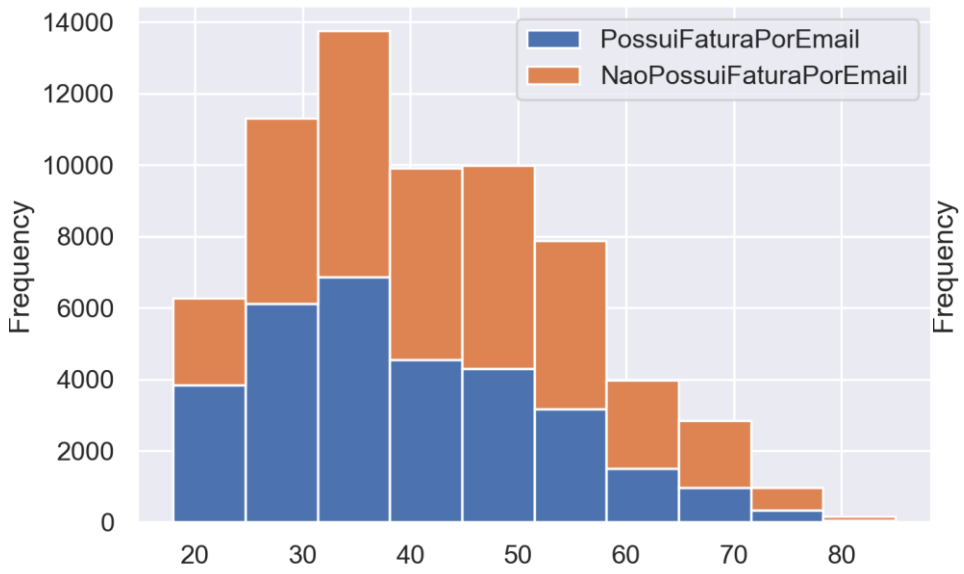
- Foi ajustado um modelo multi-classe com o classificador RandomForest com o pacote *scikit-learn* no Python;
- As colunas relacionadas as classes são: **PossuiFaturaPorEmail** e **PossuiSMSAlerta**;
- Foi usada a função sklearn.feature_selection.SelectFromModel para selecionar as características com mais importância;
- Foram selecionadas as características em que somavam mais de 85% de importância para o modelo ajustado;
- Características finais retornadas foram: **TempoNaFortbrasilEmMeses, Idade, QtdComprasUlt12Meses, QtdComprasUlt6Meses, QtdComprasUlt3Meses, QtdComprasUltMes, Cidade, Atividade_Emissor** e **PossuiEmailCadastrado**.

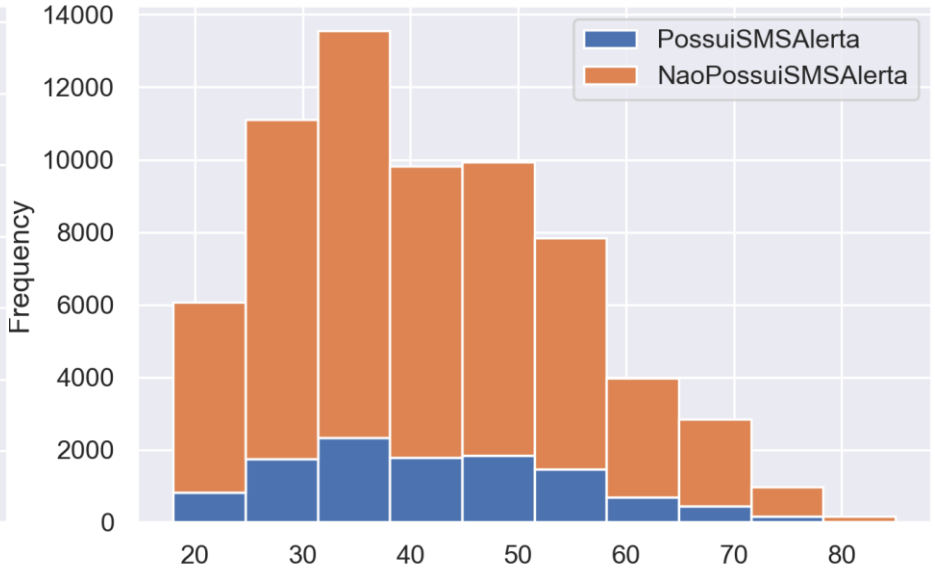
Histogramas para tempo na FortBrasil em meses



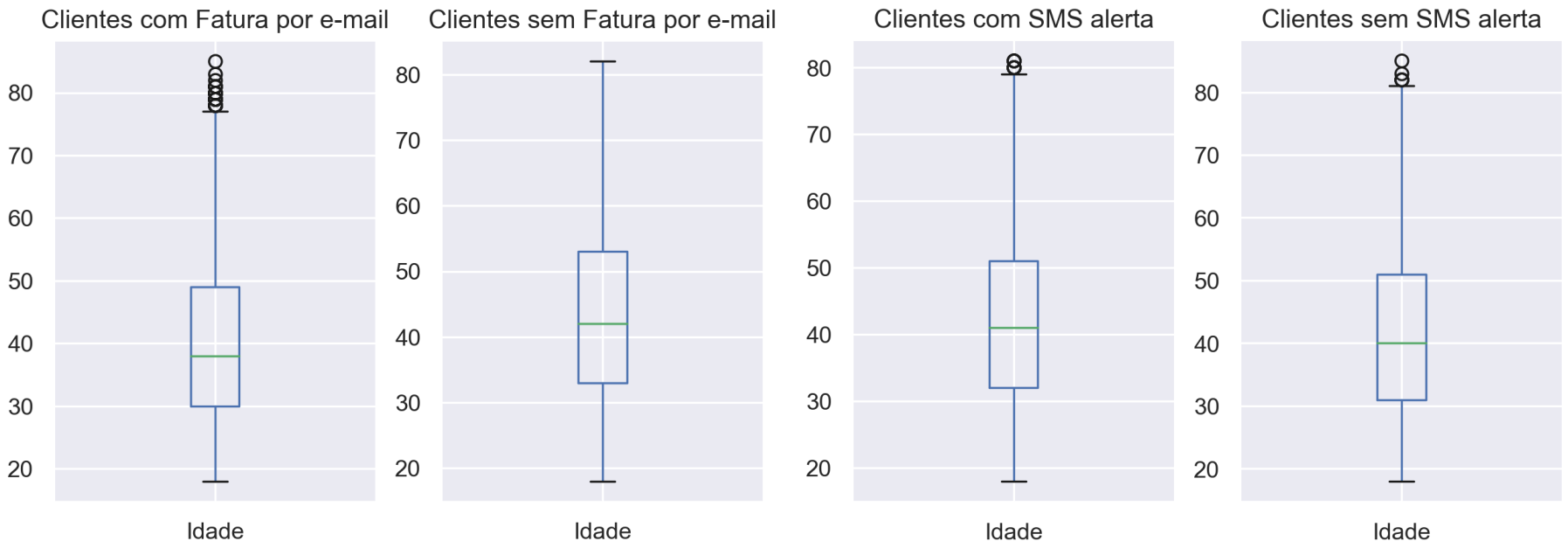


Histogramas para idade dos clientes que possuem fatura por e-mail ou SMS alerta



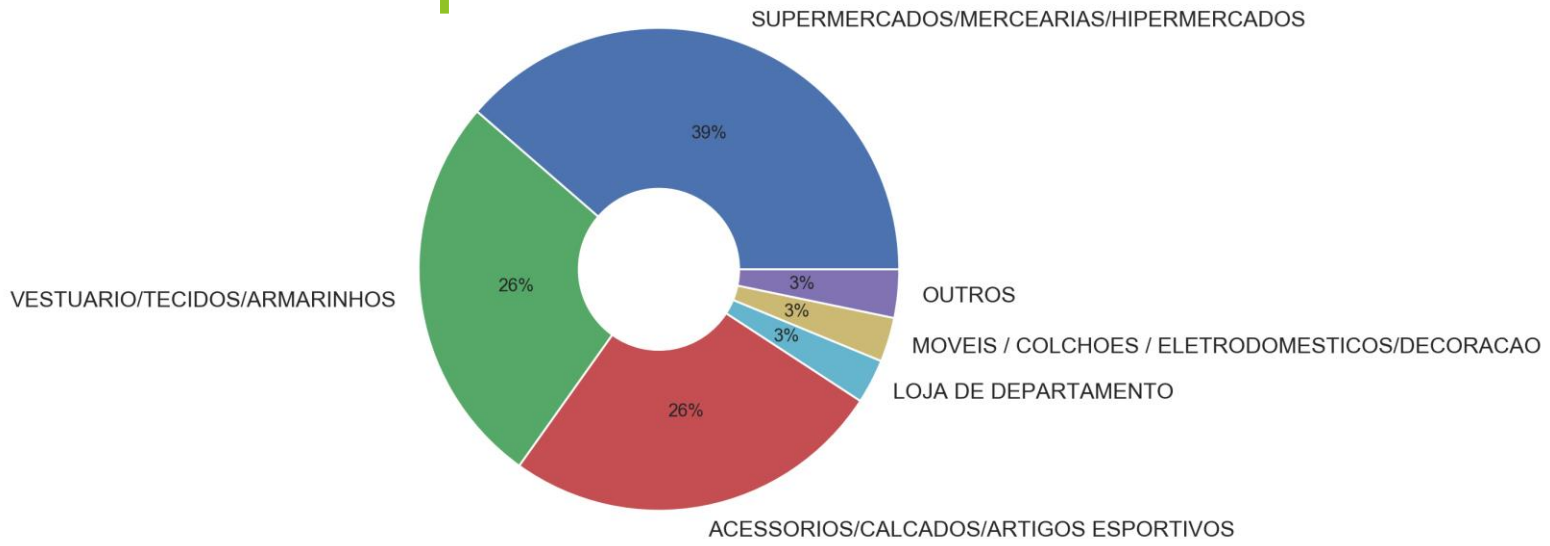


Variação de idade dos clientes que possuem, ou não, fatura por e-mail e SMS alerta

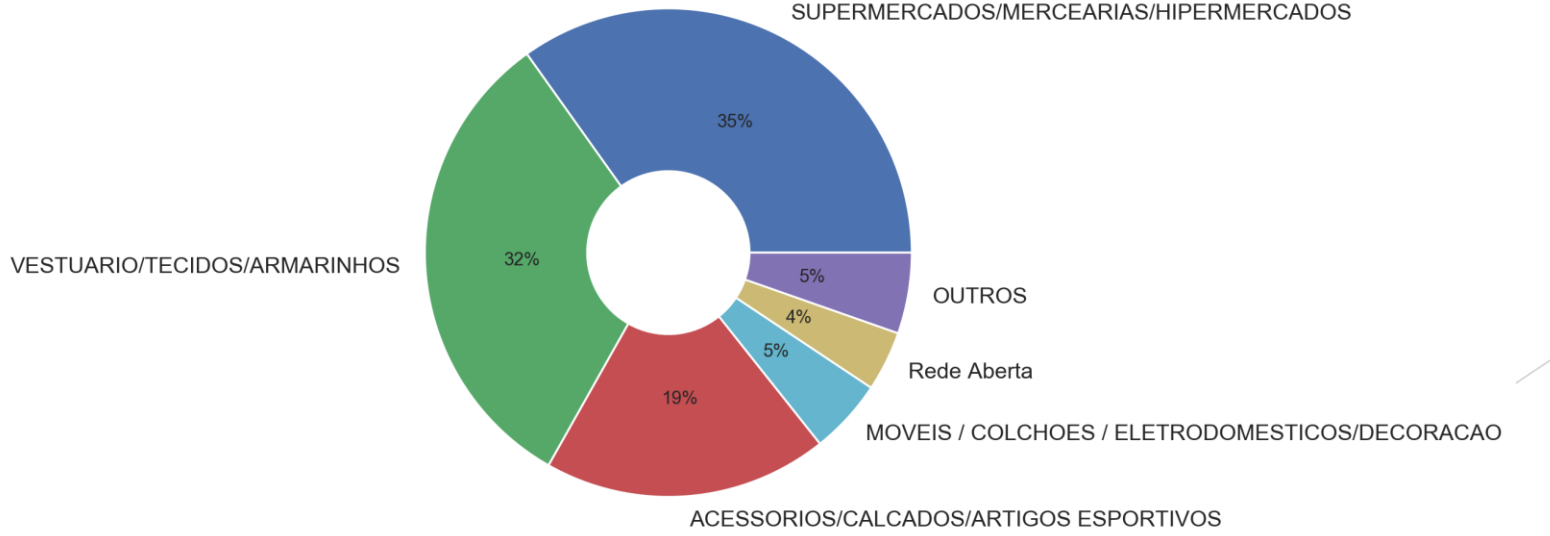


Clientes que possuem fatura por e-mail ou SMS Alerta por setor

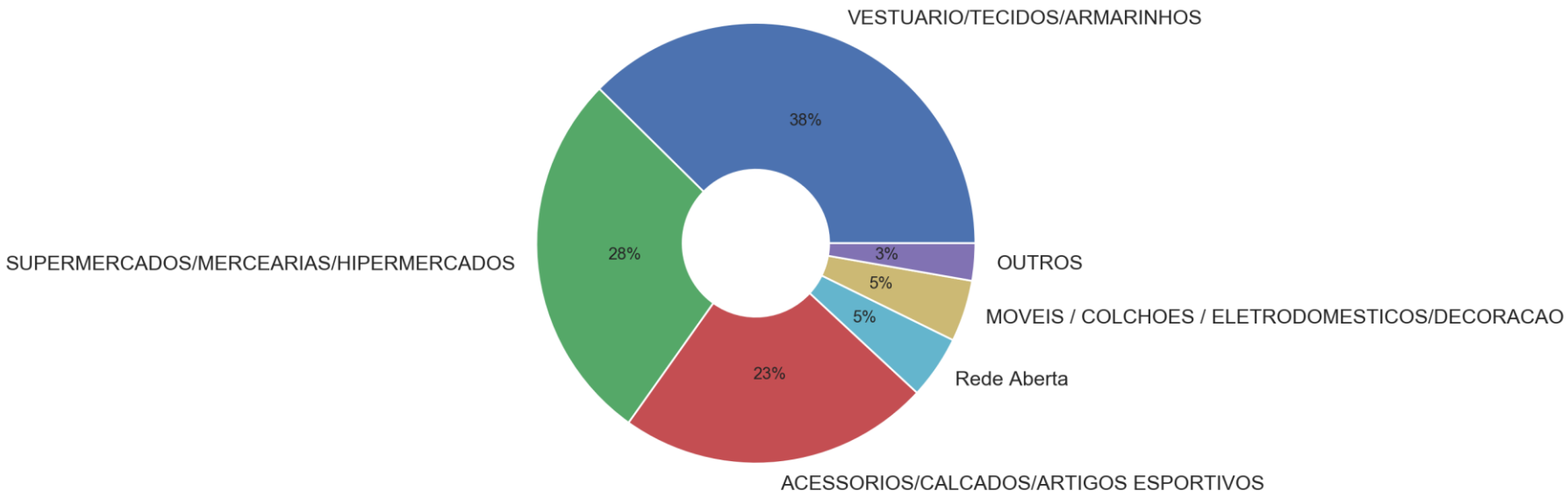
Com fatura por e-mail



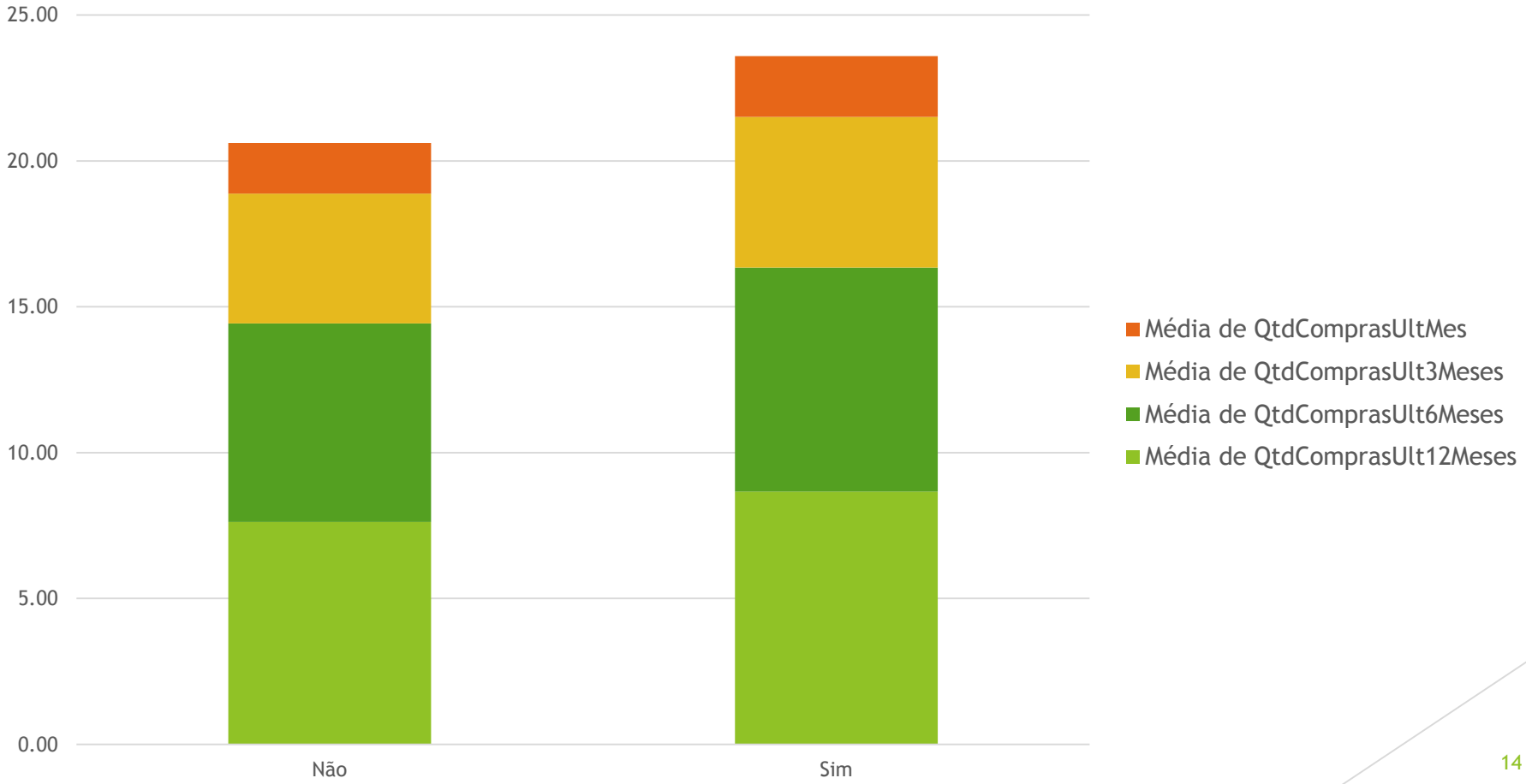
Com alerta SMS



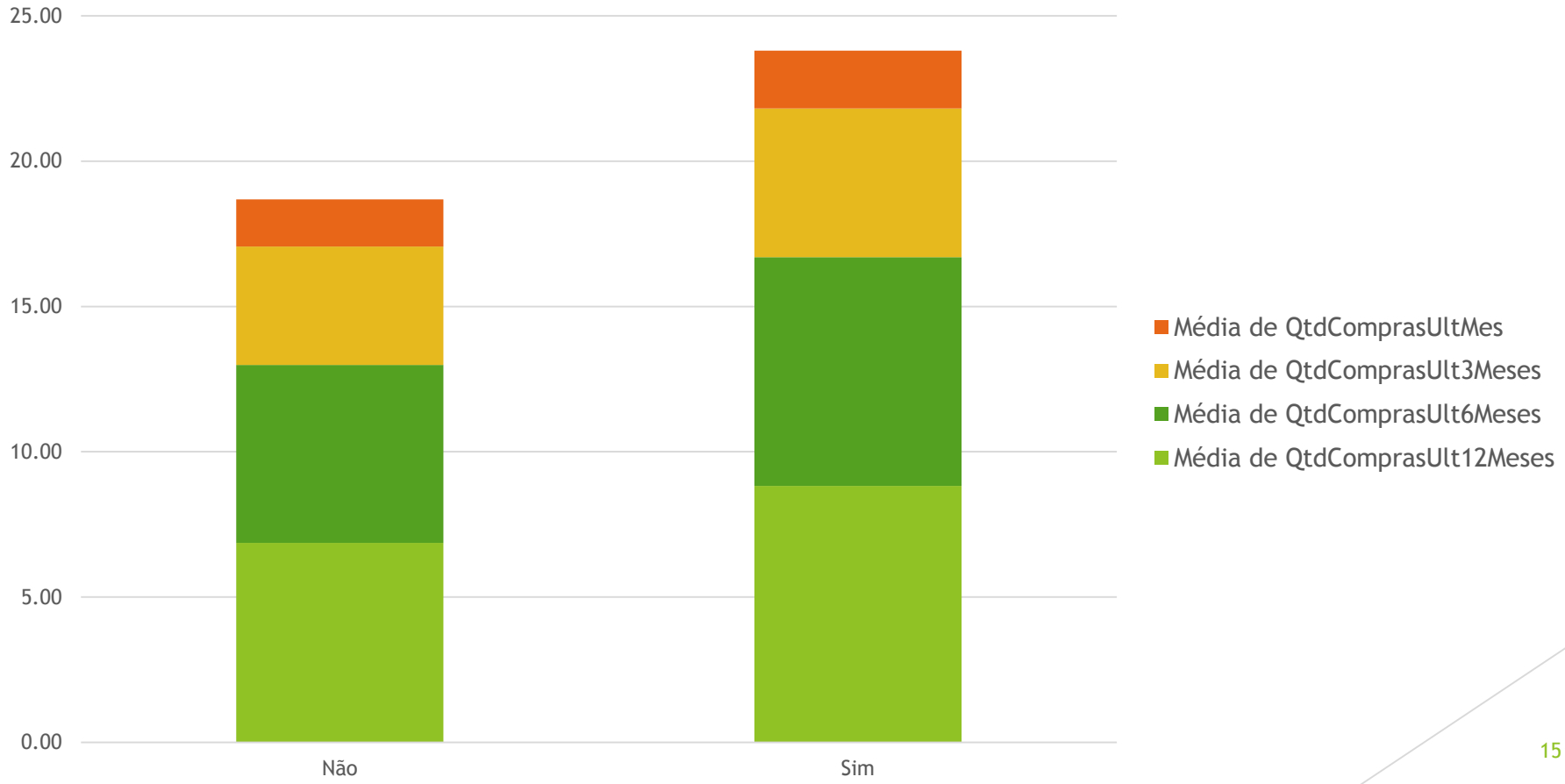
Clientes que não possuem ambos os produtos por setor



Clientes que possuem, ou não, SMS Alerta



Clientes que possuem, ou não, fatura por e-mail

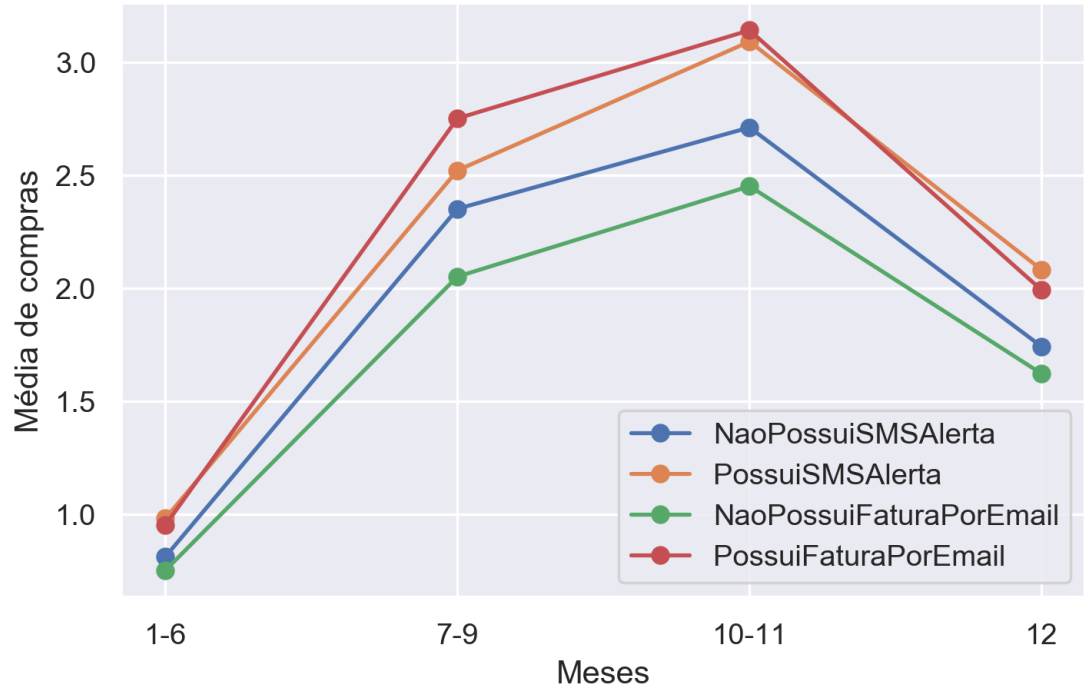


Médias de compras dos últimos meses

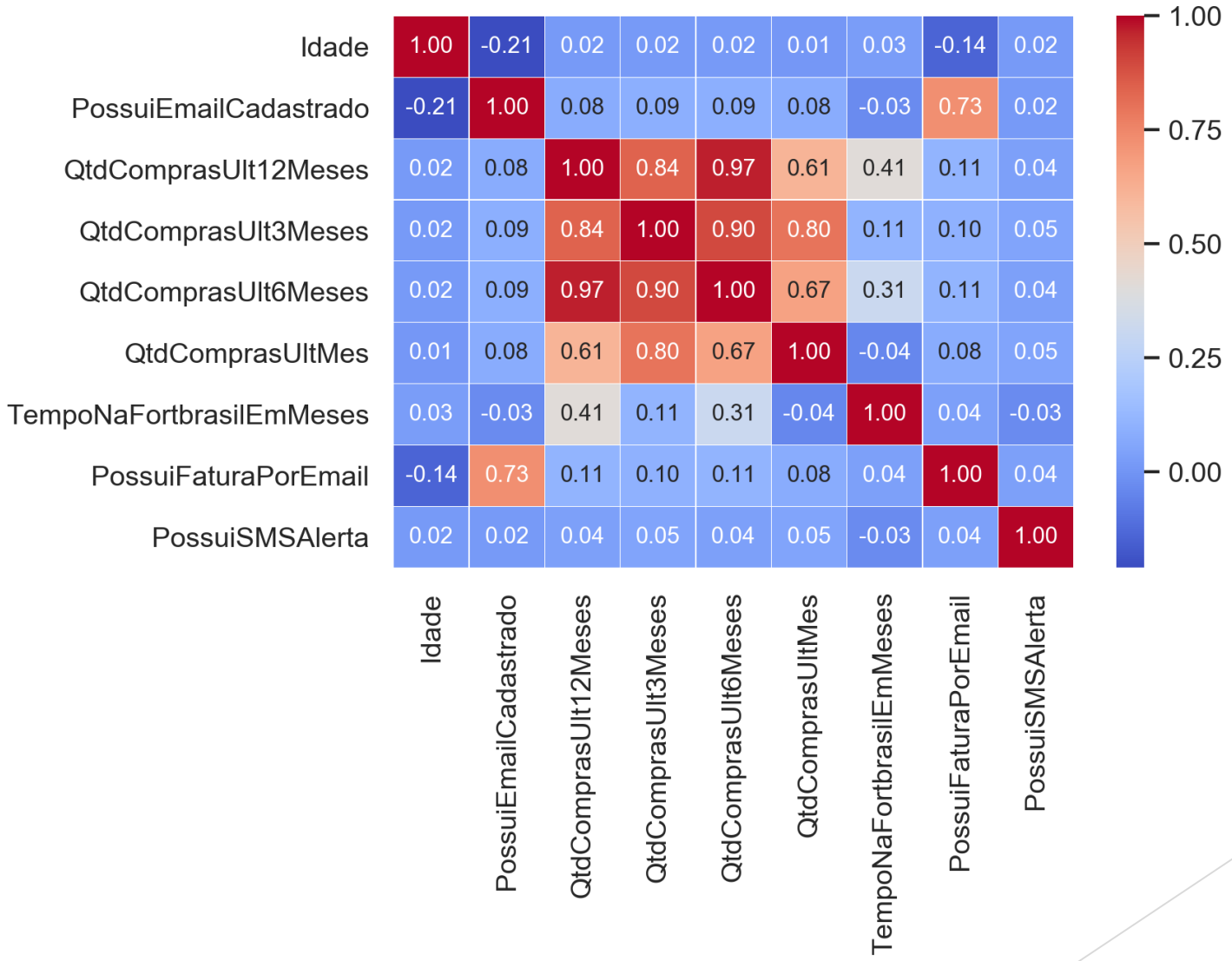
PossuiSMSAlerta	Média de QtdComprasUlt12Meses	Média de QtdComprasUlt6Meses	Média de QtdComprasUlt3Meses	Média de QtdComprasUltMes
Não	7.62	6.81	4.45	1.74
Sim	8.66	7.68	5.17	2.08
Total Geral	**7.80**	**6.96**	**4.57**	**1.80**

PossuiFaturaPorEmail	Média de QtdComprasUlt12Meses	Média de QtdComprasUlt6Meses	Média de QtdComprasUlt3Meses	Média de QtdComprasUltMes
Não	6.87	6.12	4.07	1.62
Sim	8.82	7.87	5.12	1.99
Total Geral	**7.80**	**6.96**	**4.57**	**1.80**

Médias de compras ao longo dos últimos 12 meses



Correlações entre variáveis



Selecionando 1000 clientes para ofertar fatura por email

- Tratamento de variáveis categóricas:
  - Colunas relacionadas às variáveis categóricas foram transformadas variáveis *dummy*;
  - Exemplo:

Sexo	Sexo_M	Sexo_F
NaN	0	0
M	1	0
F	0	1

- Separação dos dados nos *clusters* com PossuiFaturaPorEmail = 1 (grupo 1) e PossuiFaturaPorEmail = 0 (grupo 2);
- Para o grupo 1, foi calculado um centroide, isto é, as médias de cada variável para o grupo;
- Os 1000 clientes do grupo 2 mais próximos do centroide do grupo 1 considerando distância euclidiana, foram selecionados para ofertar fatura por email;
- Em outras palavras, os 1000 clientes do grupo 2 mais parecidos com o grupo 1 foram selecionados.

Selecionando 1000 clientes para ofertar SMS alerta

- Mesmo procedimento de tratamento de variáveis categóricas utilizado no slide anterior.

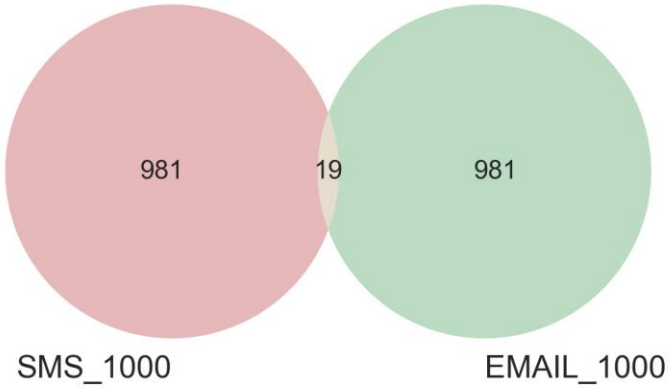
- Separação dos dados nos *clusters* com PossuiSMSAlerta = 1 (grupo 1) e PossuiSMSAlerta = 0 (grupo 2);

- O procedimento para selecionar 1000 clientes para ofertar SMS alerta foi o mesmo utilizado no slide anterior.

Clientes selecionados para oferecer cada produto

- IdConta dos 1000 clientes para ofertar fatura por e-mail no arquivo EMAIL_1000.csv.

- IdConta dos 1000 clientes para ofertar SMS alerta no arquivo SMS_1000.csv.



Tratamento dos dados para os modelos de *machine learning*

- Variáveis categóricas transformadas em variáveis *dummy, de 9* para 752 colunas após transformação;
- As colunas PossuiFaturaPorEmail e PossuiSMSAlerta foram substituídas pela coluna Classe que assume os valores:
  - Classe = 1: Ofertar SMS alerta;
  - Classe = 2: Ofertar fatura por e-mail.
- Conjunto de treinamento: clientes que possuem pelo menos um produto.
  - Clientes com SMS alerta e sem fatura por e-mail recebem classe 1;
  - Clientes com fatura por e-mail (com ou sem SMS alerta) recebem classe 2.
- Normalização de dados para seguir uma distribuição normal com média 0 e variância 1;
- Análise de componentes principais: para redução da dimensão dos dados e melhor tempo de processamento dos algoritmos foi utilizado apenas 10 componentes principais.

Modelo para escolher qual produto ofertar a um dado cliente

- Modelos selecionados para testes:
  1. Máquinas de vetores suporte (SVM): kernel linear, C = 1.
  2. Máquinas de vetores suporte (SVM): kernel exponencial, $\gamma = 0.7$, C =1.
  3. Máquinas de vetores suporte (SVM): kervel polinomial de grau 3, C = 1.
  4. Árvore de decisão com hiperparâmetros padrão.
  5. Floresta Aleatória com hiperparâmetros padrão.
  6. Rede Neural: solver para otimização dos pesos ‘lbfgs’, 2 camadas de tamanhos 5 e 2, $\alpha = 1e\text{-}5$.
  7. Regressão logística com hiperparâmetros padrão.
- Para validar os modelos foi utilizada a cross-validation com avaliação de precisão e 10 dobras;

Modelo para escolher qual produto ofertar a um dado cliente

Modelo	Precisão
1	85.22%
2	94.79%
3	88.51%
4	90.65%
5	94.22%
6	93.62%
7	86.85%

- Modelo final escolhido para ajustar o conjunto de dados: Modelo 2, máquina de vetores suporte com kernel exponencial, $\gamma = 0.7$ e $C = 1$.

Matriz de confusão para modelo ajustado com o conjunto de dados

		Classe prevista	
		1	2
Classe real	1	4055	1402
	2	297	31167

Exemplo de predição

TempoNaFortbrasilEmMeses	Idade	QtdComprasUlt12Meses	QtdComprasUlt6Meses	QtdComprasUlt3Meses	QtdComprasUltMes	Cidade	Atividade_Emissor	PossuiEmailCadastrado
11	53	6	3	1	1	FORTALEZA	Rede Aberta	0

- Transformado em variáveis *dummy*;
- Dados normalizados;
- Reduzido a 10 componentes principais;
- Predição do modelo ajustado: Classe 1, ou seja, oferte SMS alerta a esse cliente.

Previsão do conjunto de testes

- Para o conjunto de testes foi utilizado o grupo de clientes que não tinham ambos os produtos.