Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Aphonso Henrique do Amaral Rafael

Comportamento do cliente: estudo de caso de uma empresa prestadora de serviços de contabilidade digital

Curitiba
2020

Aphonso Henrique do Amaral Rafael

Comportamento do cliente: estudo de caso de uma empresa prestadora de serviços de contabilidade digital

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Fernando de Pol Mayer

Curitiba
2020

Comportamento do cliente: estudo de caso de uma empresa prestadora de serviços de contabilidade digital

Aphonso H. A. Rafael[1]

[1]Departamento de Estatística, Universidade Federal do Paraná.

Esta pesquisa se propõe a ser um primeiro passo num projeto maior de *customer success* (sucesso do cliente), numa empresa que presta serviços de contabilidade digital, é um estudo de caso a partir do atual cenário da empresa. Acredita-se que, através de dados e técnicas de *machine learning*, podem ser extraídas informações e *insights* importantes que direcionem projetos, ações e a tomada de decisão, implicando numa melhor prestação do serviço e a melhoria da experiência do cliente. Com os resultados, a empresa ganha material para análises exploratórias e descobertas no desenvolvimento de seus objetivos. Empresas jovens, em modelos de negócio inovadores/disruptivos, carecem de modelos ou *benchmarks* de atuação, pois praticamente não há pares ou experiências similares no mercado. Através dos resultados desta pesquisa, foi possível identificar que há distinção entre os perfis de clientes, como também na forma como estes perfis interagem com a empresa. Sob este contexto, a pesquisa sugere uma metodologia para análise de dados do perfil dos clientes, bem como na forma e no porque estes interagem com a empresa, possibilitando elaborar ações, seja de processos ou tecnologia, que melhorem a experiência como um todo.
Palavras-chave: contabilidade digital, perfil de cliente, atendimento, service-desk, experiência ao cliente.

This research proposes to be a first step in a larger project of customer success, in a company that provides digital accounting services. It is a case study from its current scenario. Inside the company, staff believe that through data and machine learning techniques, they could get important insights and information that could guide projects, actions plans and support the decision making process, which might help company to provide better services and improve customer experience. From obtained results, company gets material for further exploratory analysis and discover new ways to develop its objectives. Young companies, that operate in innovative and disruptive business models, have a lack of benchmarks to follow or inspire themselves, as there are practically no peers or similar experiences in this market. Through the results of this research, it was possible to identify there are differences between customer profiles, as well in the way they interact with the company. In this context, this research suggests a methodology for analyzing customer profile data and how they interact with the company. Through this, the company is able to develop new actions, either on processes or technology, that could improve the full customer experience.
Keywords: digital accounting, customer profile, customer service, service-desk, customer experience.

1. Introdução

A contabilidade é uma ciência antiga e que permanece em constante evolução, com registros anteriores ao nascimento de Cristo até os dias atuais. A base da contabilidade moderna, utilizada em todas as empresas do mundo, é conhecida como o "Método das Partidas Dobradas" (SCHIMIDT, 2008)[1]. Negra (2003, p. 1)[2] destaca que "o Método das Partidas Dobradas foi exposto, oficialmente, por Luca Pacioli em 1494, em sua obra Suma de Aritmética, Geometria, Proporção e Proporcionalidade"; ainda que a autora também destaque e demonstre que o método é anterior a Pacioli, existindo registros históricos ainda anteriores ao século XVI (NEGRA, 2003)[2].

Como cita Imperatore (2017)[3] a contabilidade é tratada como uma ciência social, cujo objetivo é estruturar um sistema de avaliação e fornecimento de informações, a fim de "[...] prover seus usuários com relatórios, demonstrações e análises de natureza econômico-financeira da empresa em determinado momento, bem como sua evolução em determinado período" (IMPERATORE, 2017, p. 12)[3]. Ainda que evidentemente antiga, a contabilidade é um assunto técnico e pouco compreendido por quem não é, necessariamente, um estudioso da área ou frequente usuário dos seus méto-

dos, processos e relatórios.

Com a aceleração digital que o mundo vem enfrentando em todos os segmentos, os serviços contábeis também vem passando por muitas mudanças, seja na forma como é prestado, até mesmo em como o empresário lida com suas rotinas contábeis. É neste contexto que começam a surgir no Brasil, empresas especializadas na prestação deste serviço de maneira totalmente online, denominadas de "Contabilidade Digital". No Brasil, a partir de 2013 este movimento passou a ganhar força com o "[...] uso da tecnologia para automatizar processos e rotinas na regularização das contas das empresas" (ISTOÉ DINHEIRO, 2020)[4].

Não apenas no contexto de prestação de serviços contábeis de maneira digital, estas empresas também têm sido importantes responsáveis no acesso e democratização à formalização de novos negócios, antes informais (sem registro, CNPJ, etc) e também ao facilitar a abertura de novas empresas; formalizando atividades de diversos tipos de empreendedores (ISTOÉ DINHEIRO, 2020)[4]; sobretudo as de pequeno porte que, em 2008, eram mais de 10 milhões de empreendimentos no Brasil, sendo 98% deles informais (EXAME, 2008)[5].

Nesse cenário, esta pesquisa é um estudo de caso de uma das principais empresas de Contabilidade Digital em atuação no Brasil, que é um mercado é novo, tanto para os usuários, quanto para a própria empresa - que também tem poucos anos de existência e é vetor deste novo paradigma. Por isso, diversas questões precisam ser analisadas e discutidas, para que a empresa possa entregar uma experiência cada vez melhor aos clientes, assim como aprimorar seus processos internos, alocação de recursos, produto e a combinação de pessoas e tecnologia, para alavancar sua eficiência, crescimento e perenidade.

2. Contextualização e motivação

Como citado, a contabilidade é um assunto complexo e muitas vezes subjetivo até mesmo para quem é da área. Micro e pequenos empresários, usuários dos serviços de Contabilidade Digital (e novos empreendedores) não costumam dominar o tema. Por isso, acredita-se que existem diversos tipos de empresas e negócios e, para cada um deles, dúvidas e necessidades/demandas distintas em sua jornada.

Pragmaticamente, por esta pesquisa busca-se encontrar um caminho para oferecer uma proposta de valor diferenciada aos clientes da empresa, que vá ao encontro de suas necessidades, conforme a demanda que eles possuem, ou seja, num viés de personalização da prestação do serviço conforme o tipo do cliente e o que ele busca. Para tal, a pesquisa foi estruturada em 3 grandes etapas:

1. Conhecimento, por parte da empresa, de quem são seus clientes (quais são os tipos/perfis);

2. Conhecer os tipos de demanda que seus clientes possuem; isto é: quando o cliente precisa interagir com o atendimento da empresa (*service-desk*); que tipo de demandas eles buscam;

3. A partir dos resultados obtidos nos itens 1 e 2; analisar os perfis Vs. demandas para compreender as diferenças de comportamentos e necessidades dos clientes em sua jornada. Espera-se que, a partir dos resultados, a empresa possa encontrar respostas e elaborar ações, seja de melhoria de processos, como de produto ou tecnologia, para aprimorar a experiência do cliente e sua alavancagem, escala.

3. Metodologia e bases de dados

Para cada uma das etapas, foram utilizadas técnicas distintas de análise de dados, uma vez que buscavam endereçar questões diferentes. Todos os dados utilizados são da própria companhia, extraídos pelo autor dos bancos de dados, por meio de consultas em *SQL*. Uma vez que aqui são tratados dados de clientes e, em alguns casos, sigilosos, buscou-se priorizar análises agregadas e preservar a confidencialidade.

Etapa 1: a empresa possui hoje mais de 20 mil clientes, evidentemente, não há como analisá-los - ou imaginar uma prestação de serviço customizada - de forma individual, por isso, o objetivo foi agrupar (clusterizar) estes clientes e identificar os perfis, semelhanças e diferenças. Uma vez que também não se sabia o número de possíveis perfis dentre essa massa de clientes, foram utilizadas técnicas de *machine learning*, por meio de Agrupamento Hierárquico e análise de dendrograma. Após identificar o número de *clusters*, utilizou-se o *K-Means* para agrupamento e nova avaliação dos resultados e análise das variâncias explicadas a partir do números de *clusters*, pelo *Elbow Method*. As *features* consideradas para tal foram selecionadas pelo autor, conforme critérios negociais da empresa, isto é, incluindo no modelo variáveis que do ponto de vista de negócio poderiam influenciar nos tipos de demandas/necessidades dos clientes, durante sua jornada.

Etapa 2: buscou-se compreender os tipos de demandas que os clientes possuem com os serviços prestados.

Existem dois canais para atendimento (no *service-desk*) com a empresa, ambos são por escrito e inteiramente registrados: i) abertura de chamados (tickets), via troca de mensagens por email e ii) abertura de atendimento instantâneo via troca de mensagens (chat). O *service-desk* é segmentado em três níveis: 1) nível 1: topo do funil de atendimentos, isso é, a entrada de todo tipo de dúvida e de cliente. Aqui são atendidas dúvidas mais simples, de usabilidade do sistema ou rotinas básicas do cliente. 2) segundo nível, onde chegam dúvidas técnicas mais complexas, como questões legais, tributárias, ou para tipos de negócio/clientes específicos, que precisam de análise profissional detalhada e tomada de decisão. 3) o nível 3, por fim, refere-se a prestação de serviços, como alterações de documentos, reabertura de balanços, consultoria paga (legal e tributária), ajustes contábeis ou demais atividades que extrapolam os serviços básicos cobertos no contrato. Os atendimentos dos níveis Avançado e Premium (cfe. seção 4.2.1) agregam os níveis 1 e 2 numa única célula de atendimento, porém somente para os clientes que aderem a estes planos. Desta forma, foi selecionada uma base de dados contendo os atendimentos realizados nos níveis 1, 2, 3, Avançado e Premium, durante o período de sete meses (janeiro a julho de 2020). O total de itens analisados (tickets + chats) foi de aproximadamente 77 mil atendimentos e, nesta base, foram utilizadas técnicas de *machine learning* para mineração de texto (MinTex) / processamento de linguagem natural *(Natural Language Processing - NLP)*. Em sequência, foram treinados e testados modelos de classificação a partir de uma base rotulada de 383 itens - que foi classificada pela empresa para esta pesquisa - categorizando estes atendimentos entre as opções i) dúvida; ii) problema; iii) solicitação e, por fim, foram preditos/classificados os atendimentos.

Etapa 3: a partir dos resultados dos itens 1 e 2; foram realizadas análises exploratórias e estatísticas descritivas. Na sequência, os resultados dos itens 1 e 2 foram combinados para identificar macro-comportamentos, com o objetivo de prover então à empresa, os *insights* propostos nesta pesquisa. Devido ao grande volume de dados, o processamento dos algoritmos e análises foram realizados em uma nuvem fornecida pela própria empresa, em ambiente GCP (*Google Cloud Platform*). A configuração da máquina foi uma *Virtual Machine standard* do GCP (*n1-highmem-16*); cuja infraestrutura é composta de 16 vCPUs com 104 GB de memória e sistema operacional Debian. As modelagens, técnicas estatísticas e de *machine learning* foram executadas na linguagem *Python*, versão base 3.8; por meio das principais bibliotecas utilizadas para essas finalidades, tais como Scikit-Learn, SciPy, NLTK e as clássicas bibliotecas de manipulação e visualização de dados como Pandas, Numpy, Matplotlib e Seaborn. Os códigos, assim como os requerimentos com o detalhamento das bibliotecas e versões (`requirements.txt`), estão disponibilizados no *Github*[1].

4. Experimentos e análise dos resultados

4.1. Clusterização de clientes

O *dataset* com dados dos clientes foi elaborado de maneira em que cada linha representa um CNPJ (ou seja, uma empresa) e as colunas são as *features* (variáveis) para clusterização. Originalmente, ele possuía além da chave (CNPJ) outras 17 variáveis, sendo 10 numéricas e 7 categóricas. Uma vez que não é trivial realizar agrupamento por variáveis categóricas, foi necessária uma etapa de pré-processamento de dados, transformando estas variáveis categóricas em numéricas (SARKAR, 2019)[8]. Convencionalmente, duas técnicas são principalmente utilizadas para tal: i) Label Encoder e ii) One Hot Encoding (SHAIKH, 2018)[6]:

i) Label Encoder: define as variáveis categóricas em números inteiros entre 0 e N ($nclasses - 1$), onde N é o número de classes distintas. Se a classe se repete, o mesmo número inteiro é atribuído a todas as repetições. **Limitações:** como os valores das variáveis categóricas passam a ser de 0 a N, isso cria uma impressão de relacionamento e continuidade para o modelo, ou seja, que a classe 1 é menor que a 2, a 2, menor que a três e assim sucessivamente: $1 < 2 < 3 < N$ (relação ordinal).
ii) One Hot Encoding: segrega as variáveis categóricas entre seus valores, atribuindo um valor binário, ou seja: será criada uma coluna para cada classe, de cada variável categórica, é definido o valor zero ou um (0 ou 1) para essa coluna, sendo 1 verdadeiro e 0 falso. **Limitações:** como cada classe, de cada variável categórica, é transformada numa coluna na base de dados, o que pode aumentar consideravelmente o número de atributos do *dataset* (colunas na base), fazendo com que se amplie o risco de *Maldição da Dimensionalidade*.

Após testes realizados com ambos os métodos, optou-se, finalmente, pelo segundo: (*One Hot Encoding*), pois

[1] `https://github.com/aphonsoar/Data-Science-Big-Data-UFPR`

o número de colunas não aumentou de forma a impactar nos graus de liberdade ou causar Maldição da Dimensionalidade; além de que o outro método criaria a relação ordinal entre os valores das variáveis, influenciando na clusterização. Com isso, o *dataset* passou ter 49 *features*, sendo as mesmas 10 numéricas e 39 categóricas. Na figura 1 é possível verificar a composição do *dataset*. Em *vermelho* estão destacadas as variáveis categóricas e em *azul* as numéricas. Apenas uma das colunas apresentou *missing values*: [Idade_media_socios]. Estes registros eram cerca de 9% dos itens e foram ignorados na clusterização.

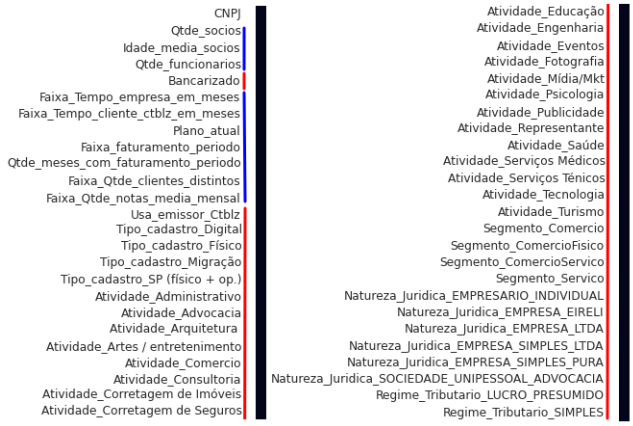

Figura 1: Dataset clientes: features consideradas

Formalmente, o método do Agrupamento Hierárquico parte da base de dados, isto é, todos os N itens do *dataset* (todos os clientes) formando N *clusters*, onde N é o número de clientes. A partir dos N *clusters*, o algoritmo encontra os 2 pontos mais próximos entre si, através do cálculo da distância euclidiana $(N_x; N_y)$ e cria um novo *cluster* agregando estes 2, num processo aglomerativo. O processo é seguido sucessivamente até que todos os valores se tornem 1 único *cluster*. Este processo é representado, portanto, pelo gráfico de dendrograma, onde é possível visualizar os níveis de cada nó da agregação e definir qual o melhor número de k, onde $k = n^o\ clusters$. Neste agrupamento foi utilizado o Método de Ward, que se consiste num "[...] procedimento de agrupamento hierárquico no qual a medida de similaridade usada para juntar agrupamentos é calculada como a soma de quadrados entre os dois agrupamentos feita sobre todas as variáveis" (SEIDEL et al., 2008, p. 10)[11]. O método é representado pela equação abaixo e utiliza um algoritmo de minimização de variâncias, que "[...] tende a resultar em agrupamentos de tamanhos aproximadamente iguais" (SEIDEL et al., 2008, p. 10)[11].

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

Na equação, u é o novo *cluster* aglomerado a partir dos *clusters* s e t. Por sua vez, v é um *cluster* não utilizado na floresta $T = |v| + |s| + |t|$, e o fator de multiplicação $|*|$ é a cardinalidade desses argumentos. Isso também é conhecido como algoritmo incremental (SCIPY, 2020)[7]. Os resultados obtidos podem ser vistos nas figuras 2 e 3 (Dendrograma e *Elbow Method Chart*).

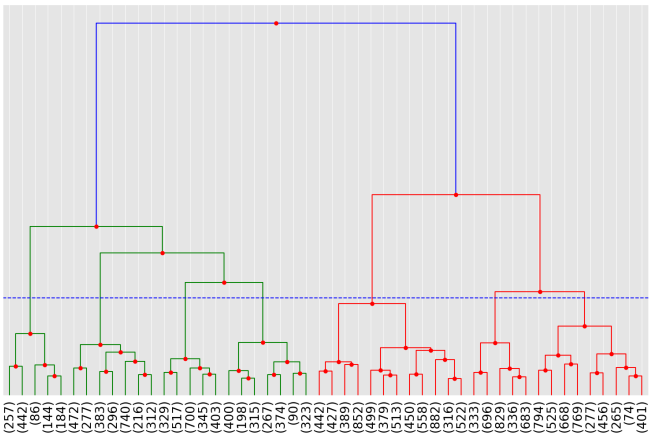

Figura 2: Dendrograma com últimos 50 nós

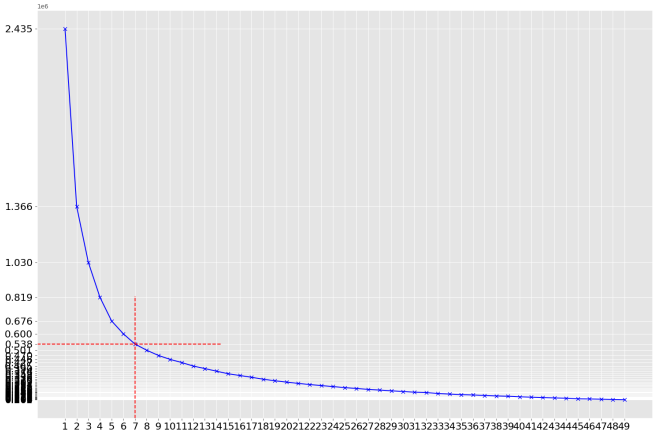

Figura 3: Gráfico do Elbow Method cortado em 7 clusters

Conforme análise do Dendrograma e também pelo ponto de inflexão do gráfico do *Elbow Method*, definiu-se portanto, a clusterização dos clientes em 7 grupos, distribuídos conforme figura 4.

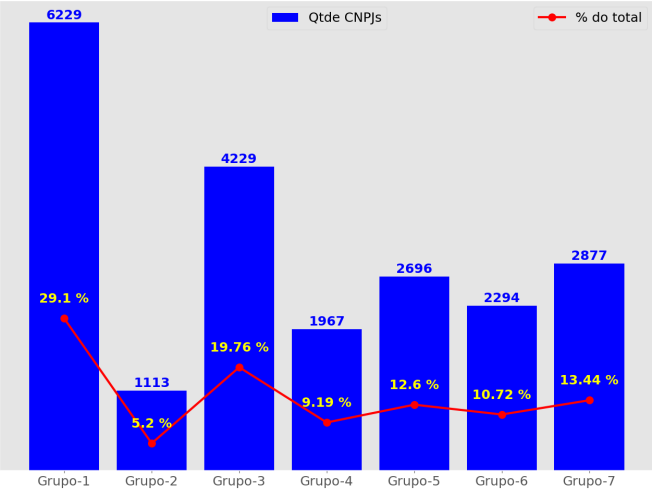

Figura 4: Distribuição dos clientes entre os clusters

4.2. Classificação e predição dos atendimentos

4.2.1. Pré-processamento e análise exploratória

Para análise dos atendimentos, foram construídos inicialmente dois *datasets*: i) tickets (emails) e ii) chats. Para ambos, todo o histórico de troca de mensagens entre o operador e o cliente foram concatenados em uma única *string* e em seguida os *datasets* foram unidos em uma base única. Quando se trabalha com *MinTex* e *NLP* também é importante realizar pré-processamento dos dados para que as *strings* / sentenças possam oferecer melhores condições preditoras (BOEHMKE, 2020) [9]. Este pré-processamento foi realizado nas seguintes etapas:

i) normalização do texto: a) palavras convertidas para *lowercase* (uma vez que os algoritmos são *case sensitive*); b) removidos os símbolos e pontuações; c) removidos espaços vazios (em branco) em excesso e; d) removidos os números (0 a 9) das *strings*.

ii) *stopwords*: são palavras que não possuem significado sintático e não colaboram com informações relevantes numa sentença, seja no sentido ou compreensão da frase, como também poderão atrapalhar ou confundir o algoritmo. São palavras como: “e”, “ou”, “para”, “de” e similares, preposições e termos de ligamento; por isso é recomendado sua remoção (SARKAR, 2019) [8]. A biblioteca NLTK possui uma lista de *stopwords* em diversos idiomas, entre elas o Português. Foi utilizada esta lista *default* com o incremento de mais algumas centenas de palavras pelo autor, conforme contexto do texto e da análise, totalizando 557 *stopwords*.

Uma vez que o volume de atendimentos analisados está na casa das dezenas de milhares, é evidente imaginar que existe uma infinidade de palavras distintas utilizadas em todas estas trocas de mensagens, por isso, inicialmente foi feita uma análise exploratória das palavras usadas, onde *insights* importantes foram extraídos.

Por meio da da biblioteca NLTK, foi utilizado um algoritmo para “tokenizar” cada palavra que aparece em cada um dos atendimentos, isto é, fazer uma uma contagem de quantas vezes cada palavra se repete em todas as sentenças. A partir disso, podemos analisar sua frequência e distribuição.

Tabela 1: Frequência das primeiras 10 palavras

	Palavra	Frequência	Percentual	Percentual acumulado
1	*empresa*	108.261	0,01133	0.01133
2	*labore*	88.762	0,00929	0.02062
3	*valor*	85.508	0,00895	0.02957
4	*contabilidade*	76.612	0,00802	0.03759
5	*pró*	69.460	0,00727	0.04487
6	*sucesso*	65.750	0,00688	0.05175
7	*plataforma*	65.461	0,00685	0.05860
8	*mês*	60.816	0,00636	0.06497
9	*pagamento*	60.387	0,00632	0.07129
10	*nota*	55.804	0,00584	0.07713

Na tabela 1, constam as primeiras 10 palavras que mais apareceram nos textos dos atendimentos, verifica-se que a palavra que mais apareceu foi “empresa” com frequência de 108.261. Sozinha, ela corresponde a mais de 1% de todas as palavras ditas. Também podemos notar que as primeiras 10 palavras que mais se repetem, representavam mais de 7% de todas as palavras escritas em todos os atendimentos. Indo um pouco mais além nesta análise exploratória, verificou-se que no total foram citadas 95.227 palavras distintas, porém, destas, apenas 5.654 já representavam mais de 95% de todas as palavras ditas. Ainda, 52.040 palavras (54% do total) apareceram somente uma única vez, ou seja, não teriam nenhum poder preditor (considerando apenas uma análise de pesos e frequências). Na figura 5, podemos ver o histograma e o percentual acumulado das primeiras 5000 palavras com maior frequência.

O histograma evidencia uma cauda muito longa, assim como sugere que não há necessidade (tampouco é recomendado) utilizar toda a base de palavras, uma vez que isso pode prejudicar o treinamento dos algoritmos com palavras com pouca ou nenhuma relevância.

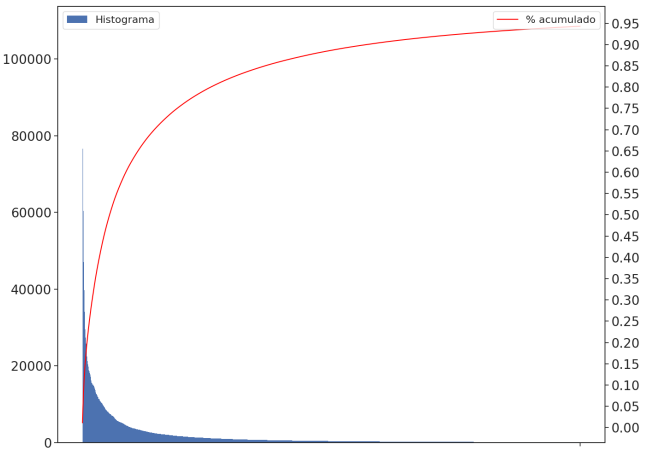

Figura 5: Histograma das primeiras 5000 palavras

Considerou-se, portanto somente as 5.654 palavras que representam 95% de toda a frequência dos textos, com objetivo de reduzir dimensionalidade e ruídos que pudessem prejudicar o modelo.

4.2.2. Classificação dos atendimentos

Realizado o pré-processamento e tratamentos, finalmente pode-se avançar para a classificação. O objetivo era compreender a natureza dos atendimentos para classificá-los entre **i) dúvida**; **ii) problema**; **iii) solicitação**. Essas três classes foram definidas pelas áreas de negócio da empresa, porém essa análise ainda nunca havia sido feita, isto é, não existia uma base rotulada de atendimentos para treinar os modelos. Portanto, a partir dos cerca de 77 mil atendimentos, foi retirada uma amostra de 383 itens para classificação: o tamanho da amostra foi calculado considerando uma margem de erro de 5% e nível de confiança de 95%. Estes 383 itens foram rotulados conforme tabela 2 e figura 6.

Tabela 2: Amostra dos atendimentos

Tipo	Frente atendimento	Qtde	Percentual	Qtde amostra
Chat	$EngajamentoN1$	31.569	41,15%	158
Ticket	$ServiçoN2$	20.893	27,25%	104
Ticket	$ServiçoN1$	14.917	19,46%	75
Ticket	$N3$	4.823	6,29%	24
Ticket	$Avançado$	2.679	3,49%	13
Ticket	$Premium$	1.811	2,36%	9

Para classificação, a amostra foi dividida numa proporção de 70% treino e 30% teste e quatro algoritmos foram executados e comparados para identificar o que apresentasse os melhores resultados. Algumas métricas de validação são comuns para analisar a performance de modelos de classificação, como i) acurácia;

Figura 6: Amostra dos atendimentos classificada

ii) valor preditivo negativo; iii) precisão; iv) *recall*; v) especificidade; vi) *F1-score* e vii) área sob a curva ROC; (SCUDILIO, 2020)[10]. Essas métricas são baseadas na matriz de confusão, a partir dos acertos, erros, falsos positivos e falsos negativos do modelo onde, dependendo do objetivo a ser alcançado, cada uma delas contribui de forma diferente na avaliação dos resultados. Uma vez que no caso em questão, buscou-se apenas identificar se o modelo previu corretamente as classes, os falsos positivos e falsos negativos possuem o mesmo peso e, por esse motivo, a métrica de acurácia (*accuracy*) é suficiente para interpretação dos resultados. A tabela 3 apresenta os resultados da acurácia dos modelos em seu estado de parâmetros *default*.

Tabela 3: Classificadores: parâmetros default

Algoritmo	Train (param default)	Test (param default)
Multinomial Naive Bayes Classifier	60,82%	45,69%
Stochastic Gradient Descent Classifier	100,00%	56,90%
C-Support Vector Classification	97,76%	47,41%
Random Forest Classifier	100,00%	50,00%

As colunas “Train” e “Test” (*param default*) representam a acurácia do modelo antes da otimização dos hiperparâmetros (*model tuning*). Para otimização, foram levantadas as possibilidades de combinação de parâmetros de cada um dos algoritmos - que fizessem sentido com o objetivo da análise - e criados *pipelines* para testes com a seleção. A partir disso, portanto, foram usados métodos computacionais para identificar a melhor combinação dos parâmetros, isto é, aqueles que permitem um melhor ajuste e generalização do modelo aos dados.

Os métodos utilizados são da Scikit-Learn e consistem em testar, exaustivamente, todas as combinações possíveis dos parâmetros definidos até que a métrica considerada como o indicador de performance (no caso a acurácia), apresente o maior valor nas bases de treino e teste. Por fim, o método também realiza a vali-

dação cruzada nas bases de treino e teste conforme a definição de *folds*. Na tabela 4, constam os os mesmos modelos após a otimização dos parâmetros e validação cruzada com 5 *folds*.

Tabela 4: Classificadores: hiperparâmetros otimizados

Algoritmo	Test (best param) cross-validation	Classificação
Multinomial Naive Bayes Classifier	60,46%	1
Stochastic Gradient Descent Classifier	63,05%	3
C-Support Vector Classification	58,95%	4
Random Forest Classifier	63,45%	2

O classificador escolhido foi o *Multinomial Naive Bayes (NB)*, pois foi o algoritmo que ficou mais equilibrado aos dados, com menor distância de acurácia entre treino e teste, o que não aparenta *overfit*, além de todas as classes terem apresentado resultado similar e satisfatório na predição da classificação (entre treino e teste). Sarkar (2019, p. 298)[8] cita que "[...] o algoritmo Naive Bayes é usado especificamente para tarefas de predição e classificação tarefas onde temos mais de duas classes" (traduzido pelo autor). Na figura 8, podemos ver a matriz de confusão do classificador:

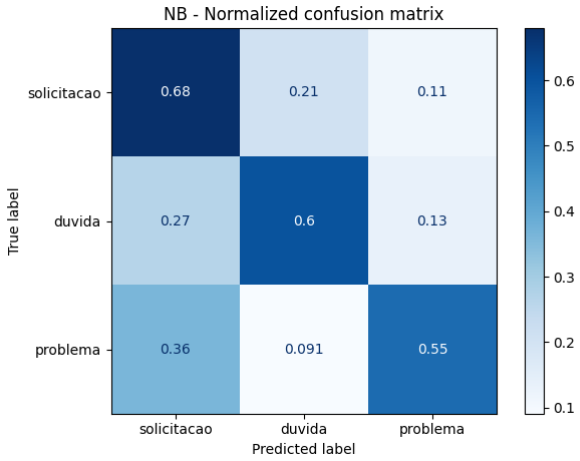

Figura 7: Matriz de confusão normalizada: Naive Bayes Classifier

Conforme figura 7, a acurácia final do classificador ficou em 60,46%. Apesar de não ser ainda um valor tão alto, esta análise não se propõe a ser exaustiva, isto é, ainda existem diversos fatores a serem aprimorados para uso na empresa; porém isso também não impede a continuidade do estudo para que seja possível extrair *insights* iniciais numa primeira observação. Dentre as ações a serem tomadas para melhoria do classificador, provavelmente a aumentar a amostra rotulada para ter uma maior base de treino seja a mais importante como *step 1*.

4.3. Combinação dos resultados da clusterização e classificação: primeiros *insights*

Combinando os resultados dos clientes clusterizados e dos atendimentos classificados, foi possível identificar comportamentos interessantes. Nas figuras 8 e 9, podemos verificar que, ainda que a distribuição dos motivos seja unânime entre os grupos - ou seja, dúvidas com maior participação, seguido por problema e solicitação - nas dúvidas, o grupo 1 se destaca de forma significativa para mais, enquanto os os grupos 2 e 5 para menos. Interessante notar que o grupo 1 também é o mais volumoso em número de clientes; ao passo em que os grupos 2 e 5 estão entre os menos volumosos (vide figura 4).

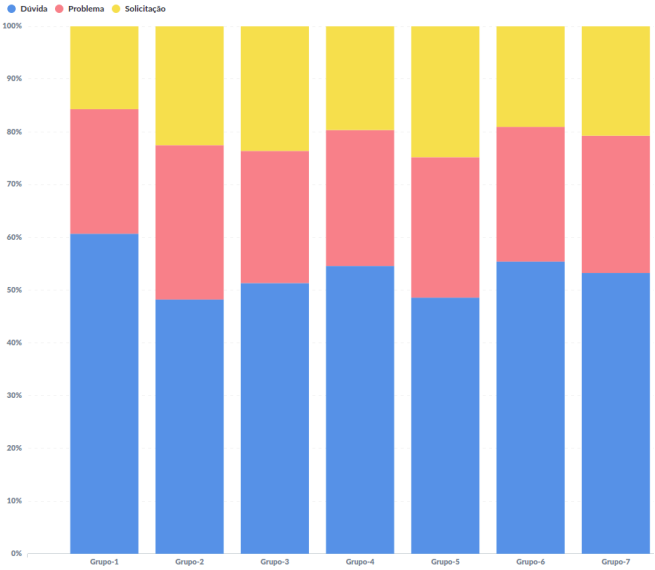

Figura 8: Distribuição das classes pelos grupos

Cluster	Dúvida	Problema	Solicitação
Grupo-1	61%	24%	16%
Grupo-2	48%	29%	23%
Grupo-3	51%	25%	24%
Grupo-4	55%	26%	20%
Grupo-5	49%	27%	25%
Grupo-6	55%	25%	19%
Grupo-7	53%	26%	21%

Figura 9: Mapa de calor das classes pelos grupos

Já na figura 10, podemos notar que os grupos 1, 4 e 6 apresentam uma taxa de contato menor do que os demais, ou seja, geram menos demanda no atendimento para a companhia. Também é interessante observar que os grupos com maiores taxas de contato (2, 3, 5 e 7) também apresentam o volume de "solicitações" maior que os demais.

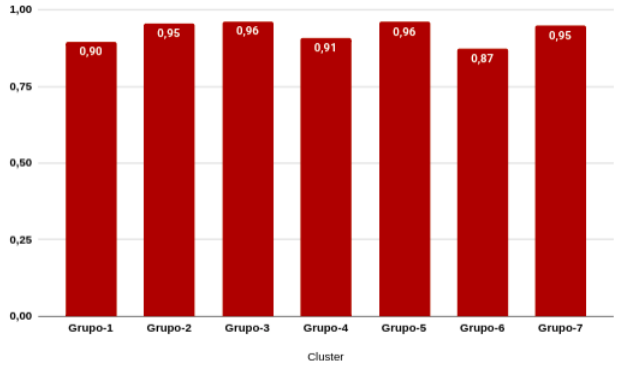

Figura 10: Taxa de contato (contact rate) por grupo

5. Conclusão

Com a análise dos resultados, realizada no item 4.3; fica evidente que existe distinção entre os tipos de clientes da empresa e na forma como estes interagem os com ela, o que era uma das questões centrais a serem respondidas por meio deste estudo. Por meio dos resultados da etapa 1 (clusterização - 4.1), foi possível verificar que a maioria das variáveis consideradas apresentaram comportamentos diferentes para cada um dos 7 grupos, o que implica na inferência de que foram úteis na separação desses clientes, permitindo compreender o que é relevante nessa separação, bem como identificar e traçar perfis para os grupos e os clientes que fazem parte deles. Já na etapa 4.2.1, a partir da classificação dos atendimentos, verificou-se que as dúvidas são o maior motivo de contato dos clientes com a empresa, sendo cerca de metade deles. A outra parte, se divide também de forma similar entre as duas demais categorias (problema e solicitação). Ou seja, fica evidente que para reduzir o volume de atendimentos demandados, o foco inicial da empresa deve ser na redução das dúvidas. Importante considerar, no entanto, que essa proporção varia entre os grupos, o que também era uma questão central a ser compreendida. Essa interpretação sugere que os tipos de clientes (perfis) possuem demandas diferentes e, desta forma, ações específicas podem ser necessárias para cada grupo. A pesquisa, no entanto, não é exaustiva e não se propôs a esgotar o assunto, mas sim em dar um primeiro passo, obter *insights* e identificar critérios que permitissem compreender os perfis de clientes e a relação destes com os tipos de demandas que fazem no atendimento *(service-desk)*. Isto é, criar uma metodologia inicial de análise para que a empresa possa entender a jornada de seus clientes e elaborar ações, seja de processos ou tecnologia, que melhorem a experiência. A partir dos resultados, cabe como próximos passos uma extensa análise exploratória, considerando as demais variáveis utilizadas na clusterização dos clientes em conjunto com as áreas de negócio da empresa, para melhor a interpretação e assim enfim, colaborar na nos planos de ação a serem formulados e no desenvolvimento do negócio, seja na experiência do cliente, ou na melhora de eficiência da companhia, via alavancagem operacional do seu time de atendimento, com automação de rotinas repetitivas ou que agregam pouco valor a companhia e ao cliente.

Referências

[1] SCHMIDT, Jose Luiz dos Santos e Paulo. *História da Contabilidade: foco na evolução das escolas do pensamento contábil*. Santos: Atlas, 2008. 176 p.

[2] NEGRA, Elizabete Marinho Serra. *Evidências das partidas dobradas através da Matemática na Mesopotâmia*. Revista do CRC/PR, Curitiba, 2003.

[3] IMPERATORE, Simone Loureiro Brum. *Fundamentos da Contabilidade*. Curitiba: Intersaberes, 2017. 173 p.

[4] ISTOÉ DINHEIRO. *Tecnologia - Contabilidade Digital: escritórios contábeis on-line ganham espaço com modelo de startup e promessa de custos menores*. 2020.

[5] EXAME. *Informalidade atinge 98% das pequenas empresas*. 2008.

[6] SHAIKH, Raheel. *Choosing the right Encoding method-Label vs OneHot Encoder*. Towards Data Science, 2018.

[7] Scipy. *Cluster Hierarchy Linkage*. Scipy Python Library, 2020.

[8] SARKAR, Dipanjan. *Text Analytics with Python: a practitioner's guide to natural language processing*. 2. ed. Bangalore, Karnataka, India: Apress, 2019. 668 p.

[9] BOEHMKE, Bradley. *Business Analytics R Programming Guide: creating text features with bag-of-words, n-grams, parts-of-speach and more*. University of Cincinnati, 2020.

[10] SCUDILIO, Juliana. *Qual a melhor métrica para avaliar os modelos de Machine Learning?* Flai Inteligência Artificial, 2020.

[11] SEIDEL, Enio Júnior et al. *Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite*. Santa Maria, 2008.