UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

VERA LUCIA VASILEVSKI DOS SANTOS ARAÚJO

CIÊNCIA DE DADOS APLICADA À AQUISIÇÃO DA LÍNGUA ORAL

Santa Helena, Paraná

2023

VERA LUCIA VASILEVSKI DOS SANTOS ARAÚJO

CIÊNCIA DE DADOS APLICADA À AQUISIÇÃO DA LÍNGUA ORAL

Data Science applied to oral language acquisition

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação da Universidade Tecnológica Federal do Paraná, Câmpus Santa Helena, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Agnaldo da Costa

Santa Helena, Paraná

2023

VERA LUCIA VASILEVSKI DOS SANTOS ARAÚJO

CIÊNCIA DE DADOS APLICADA À AQUISIÇÃO DA LÍNGUA ORAL

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 19/junho/2023

_____
Agnaldo da Costa
Doutorado
Universidade Tecnológica Federal do Paraná

_____
Arlete Teresinha Beuren
Doutorado
Universidade Tecnológica Federal do Paraná

_____
Prof. Dr. Davi Marcondes Rocha
Doutorado
Universidade Tecnológica Federal do Paraná

Santa Helena, Paraná
2023

A forças, pessoas e fatos que não me deixaram desistir,
ao me mostrar que isso não era uma opção.

AGRADECIMENTOS

A meus colegas de curso agradeço a união na trajetória;

A meu orientador, Agnaldo, a disposição;

Aos professores da banca, Arlete e Davi, a participação;

Aos emissores que realizaram os testes, os ensinamentos e o sorriso;

Aos professores aplicadores, a empolgação;

À minha família, os cafés com conversa, que ajudam a manter o ânimo e a criatividade;

A meus amigos, os vinhos e chopes com risadas, tão importantes para continuar.

"Nada neste mundo pode substituir a **persistência**.
O talento não pode;
nada é mais comum do que homens talentosos e fracassados.
A genialidade não pode;
gênios não recompensados estão em todo lugar.
A educação não pode;
o mundo está cheio de gente instruída que não chegou a lugar algum.
Persistência e determinação, apenas elas, são onipotentes".

(Calvin Coolidge)

RESUMO

O objetivo deste trabalho é auxiliar na compreensão do processo de aquisição da linguagem verbal oral, a partir da exploração de uma base de dados de fala, utilizando princípios, técnicas e ferramentas da Ciência de Dados. Um banco de dados linguísticos, composto por resultados de testes de produção oral aplicados a crianças de 3 a 7 anos, foi preparado e submetido a técnicas de Data Warehouse, ferramentas de Business Intelligence e mineração de dados. Foram aplicados K-Means, Apriori e Naïve Bayes. Dentre os resultados obtidos, destaca que: 1) a variação sociolinguística marcante da região pesquisada está sendo menos realizada pelas novas gerações; 2) por volta dos 5 anos, há uma fase de transição em que ocorrem discrepâncias relevantes na aquisição do sistema fonológico pela criança; 3) no início da alfabetização persistem casos de crianças que não pronunciam todos os sons do português; foram dados os primeiros passos na direção de um modelo que relacione características determinantes da maturação fonoarticulatória. Conclui que, apesar de algumas dificuldades, é possível processar dados linguísticos de forma satisfatória na modelagem de *data warehouse*, nas ferramentas de BI e em algoritmos de mineração de dados.

Palavras-chave: *Data warehouse*. Mineração de dados. Maturação fonoarticulatória. Fonologia.

ABSTRACT

The goal of this work is to help in the understanding of the acquisition of oral verbal language process, from the exploration of a speech database, using principles, techniques and tools of Data Science. A linguistic database, composed of results of oral production tests applied to children from 3 to 7 years old, was prepared and subjected to Data Warehouse techniques, Business Intelligence tools and data mining. K-Means, Apriori and Naïve Bayes were applied. Among the results obtained, it is worth mentioning that: the marked sociolinguistic variation of the researched region is being less performed by the new generations; around the age of 5, there is a transition phase in which important discrepancies occur in the child's acquisition of the phonological system; at the beginning of literacy, there are cases of children who do not pronounce all the sounds of Portuguese; the first steps were taken towards a model that relates determinant characteristics of phonoarticulatory maturation. It concludes that, despite some difficulties, it is possible to process linguistic data satisfactorily in data warehouse modeling, in BI tools and in data mining algorithms.

Keywords: Data warehouse. Data mining. Phonoarticulatory maturation. Phonology.

LISTA DE ILUSTRAÇÕES

FIGURAS

QUADROS

SUMÁRIO

1 INTRODUÇÃO

Para haver conhecimento, são necessários dados, fatos do mundo real. Os dados permitem conhecer uma situação, e então entendê-la. Assim, uma quantidade de dados expressiva, após processada, analisada e interpretada, pode jogar luz sobre qualquer problema complexo, permitindo conhecê-lo em detalhes, abrir caminhos para sua solução e para prevenir sua ocorrência. Isso se aplica a todas as áreas do conhecimento.

Sabe-se que, tradicionalmente, ao longo da história, a transformação de dados em conhecimento depende da análise e interpretação manuais, o que torna o processo lento, subjetivo e bastante limitado. Nas últimas décadas, no entanto, a coleta de grandes volumes de dados tem sido facilitada por recursos computacionais cada vez mais poderosos, contudo, conjuntos de dados em forma bruta são de valor restrito. Afinal, o que realmente vale é o conhecimento que pode ser extraído dos dados e colocado em uso para auxiliar de alguma forma a sociedade contemporânea. O rol de ferramentas disponíveis atualmente chegou a certo estágio em que é possível reunir informações e enxergar o que sempre esteve presente, mas nunca foi notado.

Ainda, novas abordagens podem ser vislumbradas em cima de dados conhecidos, bem como situações novas e inesperadas podem emergir na análise automatizada dos dados. Isso se justifica, porque o verdadeiro valor desses dados reside na capacidade humana de extrair relatórios úteis, identificar eventos pertinentes e tendências, apoiar decisões e respaldar políticas baseadas em análise estatística e inferência, além de explorar os dados para atingir objetivos operacionais e científicos, dentre outros (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Essas novas possibilidades propiciaram o surgimento da Ciência de Dados, que é uma abordagem multidisciplinar desenvolvida para encontrar, extrair e revelar padrões em dados por meio de uma fusão de métodos analíticos, experiência de domínio e tecnologia. Tal abordagem geralmente inclui mineração de dados, previsão, aprendizado de máquina, análise preditiva, estatística e análise de texto (DAS, 2016).

Apesar do avanço computacional das últimas décadas, a capacidade humana de analisar e entender massivos conjuntos de dados é bastante inferior à capacidade

tecnológica de reuni-los e armazená-los (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). Desse modo, o trabalho com expressivos bancos de dados demanda conhecimentos em computação – em que se destaca a área de bancos de dados, tendo em vista a tecnologia de Data Warehouse –, matemática e da área específica que se deseja pesquisar.

Nesse sentido, este trabalho aplica a Ciência de Dados à área da Linguística, mais especificamente à aquisição da língua materna, a fim de descobrir características, padrões e tendências que norteiam esse processo.

A literatura mostra que a audição é pré-requisito para a fala. Desse modo, o baixo desempenho em alguma habilidade auditiva, causado por distúrbios auditivos e de processamento auditivo, interfere diretamente na aquisição da fala e, consequentemente, na aprendizagem de maneira geral. Já a fala é pré-requisito para a alfabetização (leitura e escrita), de modo que, antes de alfabetizar, é importante conhecer a condição fonoarticulatória do aprendiz, ou seja, quais sons pertencentes ao sistema fonológico de sua língua materna ele consegue articular, para que consiga acompanhar o ensino da língua escrita (SCLIAR-CABRAL, 2003a).

Geralmente, então, é a escola – mais especificamente durante os primeiros anos – o ambiente em que essas duas situações (audição e fala) se destacam, por serem igualmente importantes à alfabetização. Casos de rendimento escolar inferior ao esperado da criança podem estar ligados a esses fatores, mas nem sempre se faz relação entre rendimento e problemas de fala – que se estendem a desvios fonológicos, decorrentes de alterações motoras, dentre outras causas –, por falta de ferramentas tecnológicas, de profissionais disponíveis para atender aos alunos e de conhecimento disponível e acessível a professores e até a pais, para avaliar e acompanhar o aprendiz.

Nesse sentido, uma investigação em uma base de dados de produção oral com volume suficiente para permitir aplicar algoritmos de mineração de dados pode contribuir para compreender o processo de aquisição da fala, de uma forma ainda não comum. Trata-se de conhecimento descoberto via interpretação de padrões e regras advindos dos dados após processados.

Um impedimento a esse tipo de pesquisa reside no fato de ser difícil coletar tais tipos de dados, reuni-los e prepará-los para análise, no entanto, dispõe-se de uma

base de dados dessa natureza, a qual é apropriada para esta pesquisa. Isso torna esta proposta exequível.

1.1 OBJETIVOS

Expõem-se a seguir os objetivos geral e específicos que se pretende atingir com o trabalho.

1.1.1 Geral

Auxiliar na compreensão do processo de aquisição da linguagem verbal oral, a partir da exploração de uma base de dados de fala, utilizando princípios, técnicas e ferramentas da Ciência de Dados.

1.1.2 Específicos

1) Modelar uma base de dados (*data warehouse*) para trabalho com linguagem verbal oral;
2) Buscar características e tendências de maturação fonoarticulatória;
3) Explorar relações entre a aquisição de coda silábica em |R| interna e externa e a produção do encontro consonantal perfeito com /r/;
4) Verificar a eficiência de ferramentas de BI e mineração de dados com caracteres fonológicos.

1.2 CONTRIBUIÇÕES DO TRABALHO

A análise proposta tem potencial para contribuir com uma área que não se beneficia tanto da Ciência de Dados: a aquisição da língua oral. Assim, as contribuições estendem-se à educação nas séries iniciais, no sentido de prever e evitar prejuízos futuros de aprendizagem, bem como entender o processo para tratar intercorrências com mais efetividade. Ainda em termos de desenvolvimento da linguagem, a Fonoaudiologia e a Linguística se beneficiam dos resultados, que

permitem conhecer fatos do desenvolvimento da linguagem e assim, respectivamente, agilizar um diagnóstico e aumentar ou aprofundar conhecimento linguístico. Ressalta-se que análises dessa natureza, com essa temática de estudo e enfoque, como exposto na subseção Estado da Arte, não são comuns, o que confere a este trabalho certo pioneirismo e gera expectativas favoráveis de utilidade.

1.3 JUSTIFICATIVA

Este trabalho insere-se nas áreas temáticas de Banco de Dados e Metodologia e Técnicas da Computação. As disciplinas que mais auxiliam em seu desenvolvimento são: Fundamentos de Bancos de Dados, Estatística, Linguagem de Programação Estruturada, Sistemas Gerenciadores de Banco de Dados, Linguagem Orientada a Objetos, Metodologia da Pesquisa, de forma especial, Fundamentos de Sistemas Inteligentes, Fundamentos de Sistemas Inteligentes, Sistemas Inteligentes Aplicados, Data Mining e Data Warehouse. A pesquisa permite a aplicação de várias técnicas matemáticas e computacionais a dados específicos de uma base ainda muito pouco explorada, portanto, inédita, sobretudo em seu conteúdo. Adaptações de técnicas e utilização de *softwares* para análise de dados são realizadas, e, aplicadas a dados da área da Linguística, configuram contribuições tanto para a Ciência da Computação quanto para a Linguística e Educação Infantil.

1.4 DELIMITAÇÕES DO TRABALHO

O alcance das análises restringe-se ao tipo de dados constantes na base de trabalho. Assim, os achados desta pesquisa aplicam-se basicamente a crianças de 3 a 7 anos de idade, falantes nativas do português brasileiro. Apesar de haver várias teorias possíveis para contribuir na análise dos resultados encontrados, opta-se neste trabalho pelas Neurociências, por seu potencial para amparar estudos linguísticos e sua crescente evolução nos últimas tempos.

Nesse sentido, é adjunta à psicolinguística – parte da Linguística que pesquisa as conexões entre questões pertinentes ao conhecimento e uso de uma língua, tais como o processo de aquisição de linguagem e o processamento linguístico, e os

processos psicológicos que se supõe estarem a elas relacionados –, área de domínio da autora desta investigação.

2 REFERENCIAL TEÓRICO

Um referencial teórico adequado a este trabalho precisa envolver os assuntos de que trata o banco de dados a pesquisar, além da teoria sobre mineração de dados e afins. O banco de dados é composto por dados de linguagem, de modo que sua coleção segue os preceitos de áreas da Linguística, como fonologia, aquisição da linguagem materna, e das Neurociências.

2.1 NEUROCIÊNCIAS E APRENDIZAGEM

Nas últimas décadas, as neurociências têm obtido avanços no esclarecimento de como funciona a aprendizagem no cérebro, sobretudo com o auxílio de imagens de ressonância magnética (IRM) (GOSWAMI, 2008; DEHAENE, 2012; BUCHWEITZ, 2012), na qual a linguagem ganha destaque. Descobriu-se que, por volta dos 5 anos de idade, há grau de amadurecimento dos circuitos neuronais e aperfeiçoamento das conexões e atividades de regiões do córtex cerebral apropriados para capacitar as crianças a receber instrução pré-escolar (BARTOSZECK; BARTOSZECK, 2012). Desse modo, o cérebro está preparado para aprender a ler após 4 anos de idade, no entanto, tal aprendizado depende do método utilizado, o qual deve respeitar os caminhos preferidos pela estrutura cerebral para aprender.

Assim que as crianças aprendem a ler, os centros de processamento da leitura são inseridos em uma rede de áreas da linguagem no lado esquerdo do cérebro, conectadas para a linguagem falada (DEHAENE, 2012), e isso já foi demonstrado em relação ao português (BUCHWEITZ; MASON; TOMITCH; JUST, 2009). Fala e leitura estão relacionadas no cérebro, de modo que um atraso na fala sugere desenvolvimento atípico da linguagem, e é um indicador de risco de uma disfunção na leitura (SHAYWITZ; SHAYWITZ, 2008).

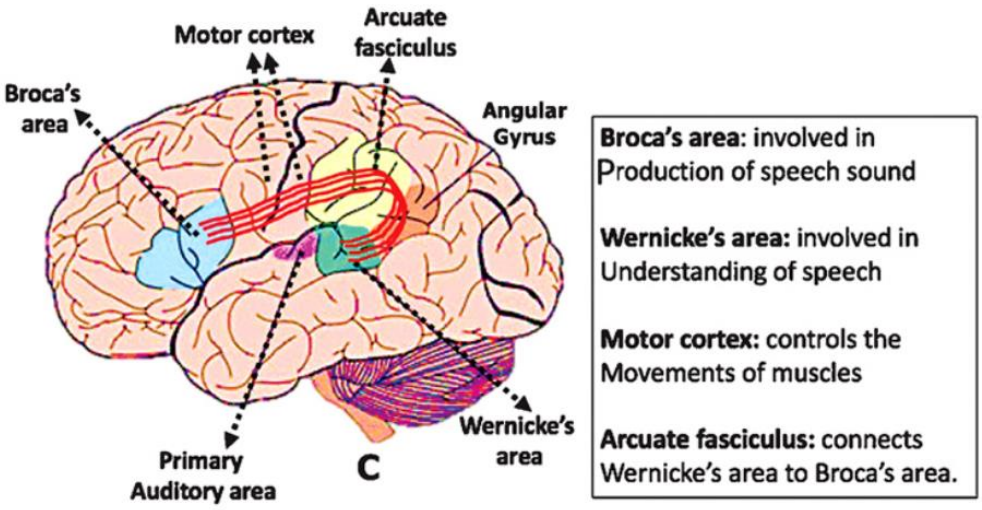
2.2 NEUROCIÊNCIAS E AQUISIÇÃO DA LINGUAGEM VERBAL ORAL

Há muitos fatores envolvidos na aquisição da linguagem verbal oral. Dentre eles, é consenso na literatura que três estão implicados no desenvolvimento da fala: inatos, maturacionais e ambientais. Fatores inatos são predeterminados pela espécie e envolvem alguns aspectos do desenvolvimento, como estrutura, amadurecimento e funcionamento do sistema nervoso central. Fatores maturacionais dizem respeito ao desenvolvimento pleno do indivíduo, isto é, quando não ocorrem adversidades que comprometam a capacidade da criança de aprender, como, por exemplo: ingestão de drogas durante a gravidez ou alimentação deficiente em proteínas durante os meses iniciais de formação da criança. A articulação dos sons da língua está relacionada aos fatores maturacionais. Fatores ambientais são os que determinam o desenvolvimento da criança de acordo com o meio em que ela está inserida. Variam entre a gestação, os primeiros meses de vida, até o processo de socialização. A especialização que os neurônios vão adquirir nas áreas secundárias do cérebro dependem diretamente da experiência (SCLIAR-CABRAL, 2003a). O fator ambiental determina qual será a língua materna.

Fisicamente, uma criança em desenvolvimento contém todos os órgãos fonadores de um adulto, diferencia-se apenas por sua posição e maturação dos tecidos celulares (WILLIAMS, 2012). Conforme a criança cresce, as posições desses órgãos são ajustadas e ocorre, naturalmente, o processo de maturação.

O cérebro da criança extrai os segmentos da fala, tria-os e os classifica. Ele explora as regularidades das cadeias que escuta para deduzir quais transições sonoras são possíveis e elimina aquelas que devem ser excluídas. A região temporal superior esquerda analisa os sons da fala, enquanto o sulco temporal superior esquerdo mostra uma organização hierárquica ligada a uma análise progressiva dos fonemas, das palavras e das frases. Já a região frontal inferior esquerda, chamada área de Broca, é tradicionalmente implicada na produção da fala e análise da gramática (DEHAENE, 2012), como ilustra a Figura 1.

Figura 1 – Partes do cérebro que controlam a fala



Fonte: Weerasooriya (2021).

Em resumo, os sons (palavras) conectam-se ao cérebro via córtex auditivo, são decodificados (compreendidos) na área de Wernicke, conectados à área de Broca pelo fascículo arqueado, na área de Broca os sons da fala são produzidos e, finalmente, no córtex motor os sons são articulados. Nesse processo, o papel do giro angular é transferir informações visuais para a área de Wernicke, a fim de facilitar o processamento de significado a partir de símbolos visuais (WEERASOORIYA, 2021).

2.3 FONÉTICA E FONOLOGIA

A ciência da linguagem ensina que as relações entre as modalidades escritas e orais de uma língua são objeto de estudo da fonologia e da fonética, respectivamente. Enquanto a fonética preocupa-se com descrever todas as línguas de forma detalhada – pois estuda os sons da fala, os fones ou alofones –, a fonologia estuda os fonemas de uma língua. Assim, a fonologia provê uma transcrição geral de uma língua, que engloba o máximo de variações possível (CÂMARA JR., 1996).

A fonologia estuda a invariabilidade profunda, enquanto a fonética abrange variações superficiais. A fonética envolve muito mais símbolos do que a fonologia, afinal, para cada variação de um fonema, há um símbolo distinto em fonética, mas em fonologia é o mesmo símbolo, pois ainda se trata do mesmo fonema, uma vez que a

variação não muda o significado da palavra em que ele é usado. Fonologia é a análise funcional do emprego que uma língua faz de seus recursos sonoros. A fonologia faz abstração das qualidades físicas dos elementos, reservadas ao domínio da fonética pura (CÂMARA JR., 1997). Os caracteres utilizados na transcrição fonético-fonológica são padronizados pelo Alfabeto Fonético Internacional (IPA, 2015), e então, possuem uma decodificação sonora única, independentemente da língua que transcrevem.

A seguir, expõe-se a representação do fonema /r/ (r fraco) e quatro de seus alofones:

fonema: /ɾ/

alofones: [ɦ], [ɣ], [ɾ], [χ]

Nesse sentido, às classes dos sons da língua dá-se o nome de fonema. Os fonemas são sons que possuem uma função determinada na língua: distinguir significados. Qualquer idioma os tem em número considerável, e as letras do alfabeto não bastam para representar os fonemas na escrita. Para remediar a deficiência, recorre-se a certas combinações, como munir algumas letras de sinais complementares (acentos gráficos nas vogais) e juntar duas letras para denotar um fonema (“lh”, “nh”, “ch” etc.) (SAID ALI, 1964).

Já a sílaba é a unidade superior, na qual os fonemas (vogais e consoantes) combinam-se para funcionar na enunciação (CÂMARA JR., 1997). A sílaba é uma divisão espontânea e profundamente examinada pela fonologia. Seus tipos de estrutura marcam caracteristicamente as línguas. A sílaba é a estrutura fonêmica elementar (JAKOBSON, 1967 apud CÂMARA JR., 1996). Na transcrição fonológica, o ponto indica fim de sílaba e o apóstrofe indica a sílaba tônica: “caixa” → /'kaj.ʃa/.

2.4 SISTEMA FONÉTICO-FONOLÓGICO DO PORTUGUÊS BRASILEIRO

O reconhecimento do sistema sonoro (fonético-fonológico) do português está diretamente relacionado à audição. Linguisticamente, o sistema fonológico refere-se ao agrupamento de fonemas de uma língua. O sistema sonoro do português engloba pelo menos 31 consoantes e 15 vogais, incluindo as vogais nasais e os ditongos. Já

o sistema fonológico do português possui 19 fonemas consonantais, 7 fonemas vocálicos orais e 5 fonemas vocálicos nasais (GUIMARÃES, 2008) – conforme Quadro 1 e Quadro 2. Alguns fonemas são simples de ser produzidos, outros possuem complexidade maior, demandando mais esforço fonoarticulatório.

Quadro 1 – Classificação dos fonemas consonantais do português brasileiro

Modo de articulação		Oclusivo			Constritivo					
					Fricativo		Lateral		Vibrante	
Papel das pregas vocais		surdo	sonoro		surdo	sonoro	surdo	sonoro	surdo	sonoro
Papel das cavidades oral e nasal		oral	oral	nasal	oral	oral	oral	oral	oral	oral
Ponto de articulação	Bilabiais	/p/	/b/	/m/	-	-	-	-	-	-
	Labiodentais	-	-	-	/f/	/v/	-	-	-	-
	Linguodentais	/t/	/d/	-	-	-	-	/l/	-	-
	Alveolares	-	-	/n/	/s/	/z/	-	-	/ʀ/	/r/
	Palatais	-	-	/ɲ/	/ʃ/	/ʒ/	-	/ʎ/	-	-
	Velares	/k/	/g/	-	-	-	-	-	-	-

Fonte: Adaptado de Vasilévski (2008) e Scliar-Cabral (2013).

Quadro 2 – Sistema de vogais fonológico do português brasileiro

	Traços	**Vogais**														
		/a/	/e/	/ɛ/	/i/	/o/	/ɔ/	/u/	/j/	/w/	/ã/	/ẽ/	/ĩ/	/õ/	/ũ/	/ɰ̃/
Função na sílaba	silábico	x	x	x	x	x	x	x			x	x	x	x	x	
	assilábico								x	x						x
Via de emissão	oral	x	x	x	x	x	x	x	x	x						
	nasal										x	x	x	x	x	x
Zona de articulação	anterior		x	x	x				x			x	x			x
	posterior	x				x	x	x								
Timbre	alto				x			x	x	X						
	médio		x			x						x		x		
	baixo	x		x			x									
Movimento dos lábios	arredondado					x	x	x						x	x	
	distenso	x	x	x	x				x	x	x	x	x			x
A semivogal /ɰ̃/ ocorre somente na palavra “muito” → /ˈmũɰ̃.tu/ e suas derivadas.																

Fonte: Adaptado de Vasilévski (2008) e Scliar-Cabral (2013).

2.5 PROCESSOS FONOLÓGICOS

Processos fonológicos são fenômenos que acontecem na pronúncia das palavras, os quais provocam alterações. Tais mudanças sonoras que ocorrem nas formas básicas dos morfemas (partes de palavras com significado, como radical e afixos) são explicadas por meio de regras que caracterizam processos fonológicos

(CAGLIARI, 2002). A maioria deles têm nomes tradicionais e é constatada em casos estudados e aceitos na literatura.

Alguns deles resultam do estado de maturação fonoarticulatória, sendo temporários e esperados, já outros são inesperados e caracterizam dificuldades que podem se prolongar, então demandam acompanhamento. Nesse sentido, destacam-se fenômenos fonológicos em sequências que exigem mais esforço articulatório, como os encontros consonantais perfeitos, ou seja, o encontro inseparável consoante + r ou l, como em “**br**aço” e “**bl**usa”, além da coda silábica em r – quando a sílaba termina em r, como em “a**r**ma**r**”.

O Quadro 3 mostra alguns dos principais processos fonológicos encontrados na fase maturacional fonoarticulatória, no qual consoantes oclusivas são chamadas de plosivas (quando há obstrução total momentânea à passagem da corrente de ar).

2.6 TESTES DE AQUISIÇÃO DA LINGUAGEM: PRODUÇÃO ORAL

Testes de aquisição da linguagem medem o quanto da capacidade esperada para determinada etapa da linguagem a criança já desenvolveu. Assim, há testes de percepção e compreensão orais, relacionados à audição; testes de produção oral, relacionados à aquisição de fonemas, mais precisamente à fonoarticulação; e testes de leitura, relacionados à compreensão (decodificação) do sistema alfabético da língua materna. Interessa aqui o teste de produção oral.

O teste de produção oral precisa ser capaz de identificar se a criança realiza os comandos dos gestos fonoarticulatórios dentro de sua variedade sociolinguística. Segundo Scliar-Cabral (2003), a criança deverá pronunciar as palavras de acordo com as imagens que o aplicador do teste apresenta, ou seja, nomeá-las. Assim, é possível verificar se a criança consegue posicionar os órgãos da fonoarticulação e realizar os gestos fonoarticulatórios corretamente, pronunciando todos os conjuntos de palavras, distinguindo entre sons, como em “maleta” → /ma.'le.ta/ e “marreta” → /ma.'ʁe.ta/.

Quadro 3 – Alguns processos fonológicos e idade esperada para superação

Processo fonológico	**Idade máxima**	**Exemplo** (forma fonológica)	**Exemplo** (forma ortográfica)
Redução da sílaba	1a6m	/sa.ˈpa.tu/ → /ˈpa.tu/	**sa**pato → pato
Harmonia consonantal	1a6m	/sa.ˈpa.tu → pa.ˈpa.tu	**s**apato → **p**apato
Plosivação de fricativa	1a6m	/ˈfa.da/ →/ˈpa.da/ /ˈsa.pu/ → /ˈta.pu/ /ˈʒa.ka/ → /ˈga.ka/	**f**ada → **p**ada **s**apo → **t**apo **j**aca → **g**aca
Simplificação de fricativa velar	3a6m	/ˈka.ʁu/ → /ˈka.u/ /ˈka.ru/ →ˈka.lu/	ca**rr**o → cao ca**r**o → ca**l**o
Posteriorização para velar	3a6m	/ta.ˈtu/ → /ka.ˈku/ /ˈdã.ma/ → /ˈgã.ma/	**t**a**t**u → **c**a**c**u **d**ama → **g**ama
Posteriorização para palatal	4a6m	/ˈsa.pu/ → /ˈʃa.pu/ /ˈze.bra/ → /ˈʒe.bra/	**s**apo → **ch**apo **z**ebra → **g**ebra
Frontalização de velar	3a	/ˈka.za/ → /ˈta.za/ /ˈgã.ma/ → /ˈdã.ma/	**c**asa → **t**asa **g**ama → **d**ama
Frontalização de palatal	4a6m	/ʃa.ˈpɛw/ → /sa.ˈpɛw /ʒa.ka.ˈrɛ/ → /za.ka.ˈrɛ/	**ch**apéu → **s**apéu **j**acaré → **z**acaré
Simplificação de líquida	3a6m	/ka.ˈre.ta/ → /kaj.ˈe.ta/ /ka.ˈre.ta/ →/ka.ˈe.ta/ /ka.ˈre.ta/ → /ka.ˈle.ta/ /ˈla.piS/ → /ˈλa.piS/ /ˈla.piS/ → /ˈa.piS/ /ˈfo.λa/ → /ˈfoj.a/ /ˈfo.λa/ → /ˈfo.la/ /ˈfo.λa/ → /ˈfo.ra/	ca**r**eta → ca**i**eta ca**r**eta → caeta careta → ca**l**eta **l**ápis → **lh**ápis **l**ápis → ápis fo**lh**a → fo**i**a fo**lh**a → fo**l**a fo**lh**a → fo**r**a
Simplificação do encontro consonantal	7a	/ˈpra.tu/ → /ˈpa.tu/ /ˈpra.tu/ → /pla.ˈto/ /ˈklu.bi/ → /ˈku.bi/ /ˈklu.bi/ → /ˈkru.bi/	p**r**ato → pato p**r**ato → p**l**ato c**l**ube → cube c**l**ube → c**r**ube
Simplificação da coda silábica	7a	/ˈpaS.ta/ → /ˈpa.ta/ /ˈpɔR.ta/ → /ˈpɔ.ta/ /a.ˈmoR/ → /a.ˈmo/	pa**s**ta → pata po**r**ta → pota amo**r** → amô
Ensurdecimento de fricativa	*	/ˈva.ka/ → /ˈfa.ka/ /ˈze.bra/ → /ˈse.bra/ /ˈʒa.ka/ → /ˈʃa.ka/	**v**aca → **f**aca **z**ebra → **s**ebra **j**aca → **ch**aca
Ensurdecimento de plosiva	*	/ˈbɔ.la/ → /ˈpɔ.la/ /ˈde.du/ → /ˈte.tu/ /ˈga.lu/ → /ˈka.lu/	**b**ola → **p**ola **d**e**d**o → **t**e**t**o **g**alo → **c**alo
Sonorização de fricativa	*	/ˈfo.tu/ → /ˈvo.tu/ /ˈsi.nu/ → /ˈzi.nu/ /ˈʃu.ti/ → /ˈʒu.ti/	**f**oto → **v**oto **s**ino → **z**ino **ch**ute → **j**ute
Sonorização de plosiva	*	/ˈpa.tu/ → /ˈba.tu/ /ˈti.a/ → /ˈdi.a/ /ˈka.za/ → /ˈga.za/	**p**ato → **b**ato **t**ia → **d**ia **c**asa → **g**asa
* Não esperado para o desenvolvimento			

Fonte: Adaptado de Wertzner (1992).

No caso, a imagem aparece na tela do computador ou celular (ver Figura 4 na seção Metodologia). Se a criança não conseguir identificar a imagem que o aplicador está apresentando, é possível que ele fale a palavra, e peça que ela a repita. Com a repetição, é possível observar se há distinção entre os gestos fonoarticulatórios e o desempenho das estruturas fonoarticulatórias. Tal teste é composto por um conjunto de palavras esperadas – nomes das figuras –, cujo objetivo é testar a produção oral de todos os fonemas do português brasileiro.

Uma possibilidade é as respostas ao teste gerarem os resultados CORRESPONDEU, NÃO_CORRESPONDEU, NÃO_RECONHECEU ou VARIAÇÃO SOCIOLINGUÍSTICA, que significam, respectivamente, que a criança pronunciou a palavra como esperado fonologicamente; que a pronunciou de forma não esperada; que não reconheceu a figura (ficou em silêncio); ou pronunciou a palavra dentro de uma variação sociolinguística. Nesse sentido, variação socolinguística é um fenômeno natural das línguas vivas, que se refere aos diferentes modos de expressar oralmente uma palavra por uma comunidade. Tais variações são também conhecidas como dialetos ou falares. Assim, podem ser representadas por processos fonológicos, posto que se referem a fenômenos da fala, porém regionalizados.

2.7 CIÊNCIA DE DADOS

Chama-se Ciência de Dados (Data Science) a área que estuda a coleta, o processamento, o tratamento, a análise, a modelagem e a visualização de dados. Trata-se de uma abordagem multidisciplinar, cuja finalidade é encontrar, extrair e revelar padrões em dados por meio de uma fusão de métodos analíticos, experiência de domínio e tecnologia. Inclui a mineração de dados, aprendizado de máquina – que se refere a como sistemas aprendem a partir de dados –, análise preditiva, estatística e análise de texto (DAS, 2016), e está estreitamente relacionada com Data Warehouse. Como a Ciência de Dados inclui recursos descritivos, diagnósticos, preditivos e prescritivos, ela viabiliza às organizações utilizar seus dados para descobrir o que ocorreu, por que ocorreu, o que ocorrerá e o que se deve esperar como resultado (MICROSOFT, 2022). A Ciência de Dados é mais do que a análise de grandes conjuntos de dados: dispõe também sobre a criação desses conjuntos. O

campo de mineração de texto expande enormemente os dados disponíveis, pois há muito mais texto sendo gerado do que números. A criação de dados de fontes variadas e sua quantificação em informações é conhecida como datificação (DAS, 2016).

2.7.1 Descoberta do Conhecimento em Base de Dados

O termo Descoberta de Conhecimento em Bases de Dados, do inglês, Knowledge Discovery in Databases (KDD) foi formalizado em 1989 – quando ocorreu o primeiro *workshop* do tema – em referência ao amplo conceito de procurar conhecimento a partir de bases de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A definição mais utilizada é a de que KDD é um processo não trivial, iterativo e interativo,[1] para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis e interpretáveis, a partir de grandes conjuntos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). A não-trivialidade do processo de KDD está na dificuldade de perceber e interpretar adequadamente inúmeros fatos observáveis na realização do processo e na dificuldade de conjugar dinamicamente tais interpretações, de forma a decidir que ações devem ser realizadas em cada caso (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

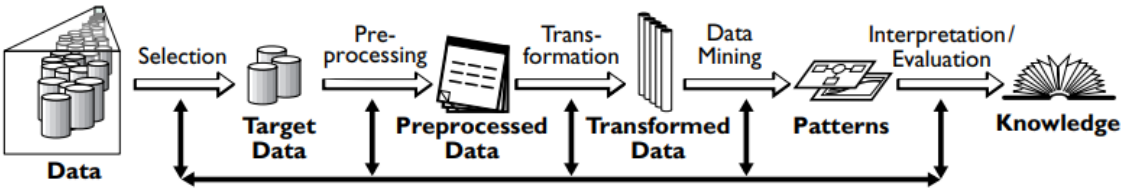
Essa extração de conhecimento envolve diversas áreas, como estatística, matemática, banco de dados, inteligência artificial, visualização de dados e reconhecimento de padrões. São utilizadas técnicas dessas áreas em diversos algoritmos de mineração de dados (CASTANHEIRA, 2008). Em se tratando de banco de dados, envolve-se a área de Data Warehouse, que guia e facilita sobremaneira a modelagem de um *dataset* para pesquisa.

Para iniciar um processo de KDD, é preciso ter claro entendimento do domínio da aplicação e dos objetivos a se alcançar. Esse processo é composto por várias etapas, que podem ser resumidas na Figura 2. Ressalta-se que, no processo de KDD, muitas decisões são tomadas pelo usuário, e pode haver *loops* entre quaisquer duas etapas (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

---

[1] **Iteração**: processo em que se procura resolver uma equação ou um conjunto de equações por meio de uma sequência de operações em que o objeto de cada operação é o resultado da operação anterior, ou seja, é um processo realizado várias vezes.
Interação: é a ação realizada pelo usuário com o produto.

Figura 2 – Visão geral das etapas de KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smith (1996).

Segundo esses autores, o pós-processamento ocorre após a etapa de mineração de dados. De acordo com vários autores (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; HAN; KAMBER; PEI, 2011; BUHA; FAMILI, 2000; FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996), as atividades de pré-processamento envolvem:

- **Criação do conjunto de dados**: inclui criar um conjunto de dados alvo, selecionar um conjunto de dados ou concentrar-se em um subconjunto de variáveis ou amostras de dados em que a descoberta será realizada;
- **Limpeza e pré-processamento**: inclui operações básicas, como remover ruídos e *outliers* – dados distorcidos ou improvisados, que geralmente ocorrem quando um sistema é desenvolvido para um propósito específico, mas é utilizado para outro (CASTANHEIRA, 2008) – ou discrepâncias, se for apropriado, coletar as informações necessárias para modelar ou levar em conta o ruído, decidir estratégias para lidar com a falta campos de dados e contabilizar informações de sequência de tempo e mudanças conhecidas, bem como decidir questões de SGBD, como tipos de dados, esquema e mapeamento de valores ausentes e desconhecidos;
- **Redução/generalização de dados e projeção**: inclui encontrar recursos úteis para representar os dados, dependendo do objetivo da tarefa, e usando redução de dimensionalidade ou métodos de transformação para reduzir o número efetivo de variáveis a considerar ou para encontrar representações invariáveis para os dados. Nessa etapa também ocorrem:

- Enriquecimento: agregar mais informação a cada registro do banco de dados para que forneça mais elementos ao KDD;
- Normalização: ajustar a escala de valores de cada atributo de forma que sejam mapeados para valores restritos a pequenos intervalos, de modo a evitar influência tendenciosa em alguns métodos de mineração de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015);
- Construção de atributos: gerar novos atributos a partir de atributos existentes (atributos derivados);
- Correção de prevalência: corrigir eventuais desequilíbrios na distribuição de certos registros, como muitos registros com a mesma característica;
- Partição do conjunto de dados: a qualidade do modelo deve ser avaliada, e uma forma para isso é utilizar dois conjuntos de dados: um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento deve conter os registros a ser utilizados na construção do modelo de conhecimento.

Em suma, pré-processam-se os dados para conhecê-los a fundo. Após isso, ocorrem as ações de mineração de dados, com:

- **Representação do conhecimento**: escolher a função de mineração de dados inclui decidir o objetivo do modelo derivado pelo algoritmo de mineração de dados (por exemplo, sumarização, classificação, regressão e agrupamento);
- **Escolha dos algoritmos de mineração de dados**: inclui selecionar métodos a ser usados para pesquisar padrões nos dados, como decidir quais modelos e parâmetros podem ser apropriados e combinar um certo método de mineração com os critérios gerais do KDD (por exemplo, o usuário pode estar mais interessado em compreender o modelo do que em sua capacidade de previsão);
- **Mineração de dados**: inclui a busca por padrões de interesse em uma forma de representação particular ou em um conjunto de tais

representações, inclui regras e árvores de classificação, regressão, agrupamento, modelagem de sequência, dependência, análise linear. Podem fazer parte também:

- Medidas de interesse: para ordenar ou filtrar os padrões descobertos ou restringir o espaço de busca da mineração de dados, a fim de melhorar sua eficiência;
- Similaridade e distância: ao interpretar o conjunto de dados como um conjunto de pontos no espaço, quanto menor for a distância entre dois pontos, maior será a similaridade entre os registros por ele representados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015);
- Aprendizado indutivo: refere-se à capacidade de determinados algoritmos de aprender a partir de exemplos. Tais algoritmos aprendem os relacionamentos entre os dados e retratam o resultado desse aprendizado nos modelos de conhecimento gerados.

Então, finalmente, há o pós-processamento, em que as partes do conhecimento extraído da etapa anterior podem ser mais profundamente processados (BUHA; FAMILI, 2000). Suas etapas podem ser:

- **Interpretação**: inclui interpretar os padrões descobertos e possivelmente retornar a qualquer uma das etapas anteriores, bem como possível visualização dos padrões extraídos, remover padrões redundantes ou irrelevantes e traduzir os padrões úteis em termos compreensíveis pelos usuários. Além disso:
  - Simplificação do modelo de conhecimento: remover detalhes do modelo de conhecimento para torná-lo menos complexo, sem perda de informação relevante. Conjuntos com muitas regras são de difícil interpretação;
  - Transformações de modelo de conhecimento: utilizar métodos de transformação sobre os modelos, para facilitar a análise. Trata-se de converter a forma de representação do conhecimento em outra forma de representação do mesmo modelo;

- Organização e apresentação de resultados: os modelos de conhecimentos podem ser representados de várias formas, como árvores, gráficos, tabelas, planilhas, cubos. São modos de visualização dos dados que estimulam a percepção e inteligência humana, aumentando a capacidade de entendimento e associação de novos padrões (GOLDSCHMIDT; PASSOS; BEZERRA, 2015);

- **Utilização do conhecimento descoberto**: inclui incorporar esse conhecimento ao sistema de origem, tomar ações com base nesse conhecimento ou simplesmente documentá-lo e reportá-lo às partes interessadas, bem como verificar e resolver potenciais conflitos com conhecimentos esperados ou previamente extraídos.

Basicamente, o pós-processamento consiste em aplicar procedimentos e métodos como filtros, interpretação, explicação, avaliação e integração de conhecimento (BUHA; FAMILI, 2000).

Por sua relevância para esta pesquisa, cabe ampliar os detalhes de algumas das atividades de KDD mencionadas, referentes ao pré-processamento e processamento dos dados.

2.7.2 Data Warehouse

A etapa da criação do conjunto de dados é respaldada pela área de Data Warehouse, que norteia e orienta a criação de um repositório de informações que possam ser analisadas para tomar decisões mais acertadas. Os dados provêm de sistemas transacionais, bancos de dados relacionais e de outras fontes, com certa periodicidade. Acessam-se os dados por meio de ferramentas de inteligência de negócios (*business intelligence* – BI), inteligência artificial, planilhas eletrônicas, dentre outras. Essa tecnologia é considerada a evolução natural do ambiente de apoio à decisão, e sua crescente utilização está relacionada à necessidade de domínio de informações estratégicas para garantir respostas e ações rápidas (MACHADO, 2013).

Assim, um *data warehouse* é um tipo de sistema de gerenciamento de dados projetado para ativar e fornecer suporte às atividades de Business Intelligence (ORACLE, 2023). Os *data warehouses* são armazéns de dados que se destinam exclusivamente a realizar consultas e análises avançadas e geralmente contêm grandes quantidades de dados históricos, portanto, são somente para leitura. Seus recursos analíticos permitem que as organizações obtenham informações úteis de seus dados para melhorar a tomada de decisões. Com o tempo, cria-se um registro histórico valioso para a análise da situação atual.

Cabe destacar que Business Intelligence é um termo utilizado para descrever um conjunto amplo, coeso e integrado de ferramentas e processos utilizados para captar, integrar, armazenar e analisar dados para a geração e apresentação de informações que deem suporte à tomada de decisões (ROB e CORONEL, 2011). Trata-se de um modelo que permite transformar informação em conhecimento.

Um *data warehouse* típico geralmente inclui os seguintes elementos (ORACLE, 2023):

- Um banco de dados relacional para armazenar e gerenciar dados;
- Uma solução de extração, carregamento e transformação (ETL – Extract, Transform, Load) para preparar os dados para análise;
- Análise estatística, relatórios e recursos de mineração de dados;
- Ferramentas de análise de clientes para visualizar e apresentar dados aos usuários de negócios;
- Outras aplicações analíticas mais sofisticadas que geram informações acionáveis, aplicando ciência de dados e algoritmos de inteligência artificial ou gráficos e recursos espaciais que permitem mais tipos de análise de dados em escala.

2.7.3 Mineração de Dados

A ação de encontrar padrões úteis nos dados é conhecida por nomes diferentes, incluindo mineração de dados, em comunidades diferentes. Por exemplo, extração, descoberta de informações, coleta de informações, processamento de padrões de dados. O termo “mineração de dados” é usado mais por estatísticos,

pesquisadores de banco de dados e, mais recentemente, sistemas de gerenciamento de informação e comunidades empresariais. Uma vez que KDD refere-se ao processo geral de descoberta de conhecimento útil a partir de dados, a mineração de dados é um passo particular nesse processo – aplicação de algoritmos específicos para extrair padrões (modelos) de dados (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Os padrões descobertos devem ser válidos para novos dados com algum grau de certeza. Também, os padrões devem ser novos – pelo menos para o sistema de origem e, de preferência, para o usuário – e potencialmente úteis para o usuário ou tarefa. Ainda, os padrões devem ser compreensíveis, se não imediatamente, após pouco pós-processamento (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Assim, mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados, para processar grandes volumes de dados. Padrões ou regras – que aqui incluem modelos ou estrutura em dados – refletem essa descoberta de informações em grandes conjuntos de dados (DAS, 2016). O volume avantajado de dados analisados contribui para as novas informações serem úteis na prática. A mineração de dados usa análise matemática para derivar padrões e tendências que existem nos dados (MICROSOFT, 2022). Como mencionado, normalmente, esses padrões não podem ser descobertos com a exploração de dados tradicional, pelo fato de as relações serem complexas ou por haver muitos dados (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). Extrair um padrão ou características também significa: ajustar um modelo aos dados; encontrar alguma estrutura nos dados; e descrever em alto nível um conjunto de dados.

Há certa variação quanto às etapas da mineração de dados, mas algumas delas podem ser: definir o problema; preparar dados; explorar os dados; criar modelos; explorar e validar modelos; implantar e atualizar modelos (MICROSOFT, 2022).

Um algoritmo na mineração de dados (ou aprendizado de máquina) é um conjunto de heurística e cálculos que cria um modelo com base nos dados. Trata-se de métodos para derivar padrões de dados (TAN; STEINBACH; KUMAR, 2009). Para criar um modelo, o algoritmo primeiro analisa os dados fornecidos, procurando tipos de padrões ou tendências específicos. O algoritmo usa os resultados dessa análise em muitas iterações para definir os parâmetros ideais para criar o modelo de

mineração. Esses parâmetros são aplicados pelo conjunto de dados inteiro para extrair padrões acionáveis e estatísticas detalhadas (MICROSOFT, 2022).

O modelo de mineração que um algoritmo cria a partir de seus dados pode assumir vários formatos, incluindo:

- Um conjunto de *clusters* que descreve como os casos em um conjunto de dados estão relacionados. Similarmente, é definido em termos de o quão próximos os objetos estão no espaço (HAN; KAMBER; PEI, 2011);
- Uma árvore de decisão que prevê um resultado e descreve como critérios diferentes afetam esse resultado;
- Um modelo matemático que faz previsões;
- Um conjunto de regras que descreve agrupamentos de produtos em uma situação e as probabilidades de que os produtos sejam pedidos juntos.

Por exemplo, o k-Means é um dos algoritmos de *clustering* mais antigos, amplamente utilizado (TAN; STEINBACH; KUMAR, 2009) e está disponível em muitas ferramentas diferentes e com implementações e opções diferentes.

2.7.4 Exploração de Dados

As técnicas de exploração incluem cálculos dos valores máximos e mínimos, cálculos das médias e dos desvios padrão e análise da distribuição dos dados. Por exemplo, ao analisar os valores máximos, mínimos e médios, é possível determinar que os dados não são representativos para o problema de estudo, e se devem obter mais dados equilibrados ou revisar as suposições que determinam as expectativas. Os desvios padrão e outros valores de distribuição podem fornecer informações úteis sobre a estabilidade e precisão dos resultados. Um desvio padrão muito grande indica que incluir mais dados pode ser útil para melhorar o modelo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Os dados que se desviam muito de uma distribuição padrão podem estar distorcidos ou representar uma imagem precisa do problema real, o que dificulta, porém, o ajuste do modelo aos dados (MICROSOFT, 2022).

Ao explorar os dados levando-se em consideração o conhecimento que o pesquisador tem sobre o problema, verifica-se se o conjunto contém dados imperfeitos. Com isso, cria-se uma estratégia para solucionar problemas e compreender comportamentos típicos. Quando se cria um modelo, criam-se automaticamente resumos estatísticos dos dados contidos no modelo, que se podem usar em relatórios ou análises adicionais (MICROSOFT, 2022).

2.7.5 Criação de Modelos (Extração de características)

Com o conhecimento obtido na exploração de dados, definem-se e criam-se os modelos, aplicando-se um algoritmo aos dados. Um modelo de mineração é um conjunto de dados, estatísticas e padrões que podem ser aplicados a novos dados para gerar previsões e fazer inferências sobre relações (HAN; KAMBER; PEI, 2011). Ao criar uma estrutura de mineração, definem-se as colunas de dados que se usarão. A estrutura de mineração é vinculada à origem dos dados, mas realmente não contém nenhum dado até que seja processada. Quando se processa a estrutura de mineração, geram-se agregações e outras informações estatísticas que podem ser usadas para análise. Essas informações podem ser usadas por qualquer modelo de mineração com base na estrutura (MICROSOFT, 2022).

Nesse sentido, cabe o discernimento de que, em qualquer conjunto de dados, podem-se encontrar padrões que parecem estatisticamente significativos, mas não são. Houve progressos substanciais na compreensão de tais questões nas estatísticas nos últimos anos (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Antes de a estrutura e o modelo serem processados, o modelo de mineração de dados apenas especifica as colunas usadas para entrada, o atributo que se está prevendo e os parâmetros que indicam ao algoritmo como os dados devem ser processados. O processamento de um modelo é chamado de treinamento. Treinamento refere-se ao processo de aplicação de um algoritmo matemático específico aos dados na estrutura com a finalidade de extrair padrões (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Os padrões localizados no processo de treinamento dependem da seleção de dados de treinamento, do algoritmo escolhido e da configuração do algoritmo.

Algoritmos diferentes são adequados a um tipo diferente de tarefa, e cada um cria modelo diferente. Parâmetros são usados para ajustar cada algoritmo e aplicar filtros aos dados de treinamento para usar apenas um subconjunto de dados, criando diferentes resultados (HAN; KAMBER; PEI, 2011). Depois de passar os dados pelo modelo, o objeto do modelo de mineração conterá resumos e padrões que poderão ser consultados ou usados para previsão.

2.7.6 Exploração e Validação de Modelos

Faz parte do processo de mineração de dados explorar os modelos de mineração criados e testar sua eficiência. Antes de implantar um modelo, ele deve ter seu desempenho avaliado por meio de testes. Além disso, normalmente criam-se vários modelos com diferentes configurações e verifica-se qual gera os melhores resultados para o problema e seus dados (HAN; KAMBER; PEI, 2011).

Usa-se o conjunto de dados de treinamento para criar um modelo e o conjunto de dados de teste para testar a precisão do modelo ao criar consultas de previsão. A partir do modelo, exploram-se tendências e padrões. Se nenhum dos modelos criados tiver desempenho esperado, redefine-se o problema ou investiga-se novamente o conjunto de dados original (MICROSOFT, 2022).

2.7.7 Implantação e Atualização de Modelos

Uma das últimas etapas do processo de mineração de dados é implantar os modelos que mostraram melhor desempenho. Após isso, é possível realizar as tarefas necessárias, como criar previsões que poderão ser usadas para tomar decisões. Com alguma ferramenta, criam-se consultas de previsão e consultas de conteúdo para recuperar estatísticas, regras ou fórmulas do modelo. A atualização dos modelos deve ser feita após a revisão e análise, e isso requer o reprocessamento dos modelos. A atualização dinâmica dos modelos, à medida que se geram mais dados, e alterações constantes para melhorar a eficiência da solução são importantes (MICROSOFT, 2022).

2.7.8 Escolha dos Algoritmos

A escolha do melhor algoritmo para uma tarefa analítica específica demanda certo esforço. Embora se possam usar algoritmos diferentes para executar a mesma tarefa, cada algoritmo produz um resultado distinto e alguns podem produzir mais de um tipo de resultado (HAN; KAMBER; PEI, 2011). Por exemplo, um algoritmo de árvores de decisão pode ser usado, além de para previsão, como uma forma de reduzir o número de colunas em um conjunto de dados, pois a árvore de decisão pode identificar colunas dispensáveis ao modelo de mineração final (MICROSOFT, 2022).

Como visto, algumas opções de algoritmos por tipo são:

- Classificação: preveem uma ou mais variáveis discretas, com base nos outros atributos do conjunto de dados; mapeiam (ou classificam) um item de dados em uma das várias classes categóricas predefinidas (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996). Tipicamente, a classificação usa aprendizado supervisionado (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A classificação tem inúmeras aplicações, incluindo detecção de fraude (HAN; KAMBER; PEI, 2011).
- Regressão: preveem uma ou mais variáveis numéricas contínuas, com base nos outros atributos do conjunto de dados. Compreendem assim a busca por funções, lineares ou não, que mapeiem os registros de conjunto de dados em valores reais (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).
- Segmentação: dividem dados em grupos ou *clusters* de itens que têm propriedades semelhantes.
- Associação: encontram correlações entre atributos diferentes em um conjunto de dados. Usados para criar regras de associação.
- Análise de sequência: resumem sequências ou episódios frequentes em dados, como uma série de cliques em um sítio da Web (MICROSOFT, 2022).

Mais de um algoritmo pode ser usado nas soluções. Analistas experientes às vezes usam um algoritmo para determinar as entradas mais efetivas (ou seja,

variáveis) e então aplicam um algoritmo diferente para prever um resultado específico baseado naqueles dados (MICROSOFT, 2022). Em uma única solução de mineração de dados, pode-se usar um algoritmo de *clustering*, um modelo de árvores de decisão e um modelo Naïve Bayes para obter exibições diferentes sobre os mesmos dados (HAN; KAMBER; PEI, 2011). Ainda, podem-se usar algoritmos diferentes em uma única solução para executar tarefas distintas (DAS, 2016), por exemplo, usar regressão para obter previsões e uma rede neural para executar uma análise de fatores que influenciam as previsões (MICROSOFT, 2022).

2.8 APRENDIZAGEM SUPERVISIONADA E NÃO SUPERVISIONADA

Os dados supervisionados e não supervisionados são dois tipos diferentes de abordagem no campo da aprendizagem de máquina e análise de dados.

Quando é não supervisionado, o aprendizado de máquina ocorre de forma independente. Com base em um número grande de dados, o algoritmo busca padrões e similaridades entre os dados, permitindo identificar grupos de itens similares ou similaridade de itens novos com grupos já definidos (MULLER; GUIDO, 2017). O algoritmo não necessita de instruções para ser executado, ou seja, com a base em dados, o próprio algoritmo se encarrega de resolver o problema, sem ação humana, apenas com a interpretação do processamento de dados.

Assim, o algoritmo não supervisionado não passa por tratamento dos dados e deve descobrir por si as relações entre os dados. Depois disso, ele verifica se há alguma equivalência com os dados para que eles sejam agrupados. Como não há dados rotulados, não é feito teste.

Para trabalhar com o aprendizado não supervisionado é necessário ter nível alto de conhecimento alto dos dados, pois, como não há interação humana, é preciso garantir que a manipulação de dados seja feita de forma correta. Caso contrário, os resultados podem não ser os esperados nem refletir a realidade, após a execução do algoritmo.

Ao contrário, algoritmos supervisionados aprendem a partir de um *dataset* de treino e devolvem os resultados com base nele, ou seja, relacionam uma saída com uma entrada, a partir de entradas e saídas conhecidas (conjunto de exemplos

rotulados). Para cada saída é atribuído um rótulo, que pode ser um valor numérico ou uma classe. O algoritmo determina uma forma de prever qual é o rótulo de saída com base em uma entrada informada (MICROSOFT, 2023a).

2.9 ESTADO DA ARTE

Tendo em vista que uma reunião de dados de aquisição da linguagem e aprendizagem da leitura em português brasileiro é incomum, não foram encontrados trabalhos similares ao que se propõe aqui. Assim, apresentam-se algumas pesquisas relacionadas e o estado de exploração da base de dados de pesquisa.

2.9.1 Trabalhos Relacionados

Uma busca por trabalhos que exploram bancos de dados fonológicos utilizando a Ciência de Dados revela que provavelmente esse tema ainda não tenha sido objeto de estudo desse tipo de análise. Há programas computacionais que leem, segmentam e preparam dados fonológicos para análise, mas os estudos não consideram Ciência de Dados.

Um deles é o Corpus CEFALA-1, uma base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia, que faz parte de um projeto desenvolvido por linguistas e um cientista da Computação (FOLLADOR NETO; SILVA; YEHIA, 2019). O foco da parte computacional é na criação do *software*, apenas. Segundo os autores, o sistema possibilita aumento significativo na coleta de dados para estudos fonoestilísticos, automatizando e padronizando vários processos manuais. A análise de tais dados parece ser puramente linguística, ou seja, contribui para a descrição do português brasileiro corrente.

2.9.2 Exploração Atual da Base de Dados

A base de dados explorada nesta pesquisa é detalhada na seção Metodologia. Aqui cabe explicar que ela foi utilizada em alguns trabalhos que avaliaram a aquisição do sistema fonoarticulatório do português brasileiro em crianças das três primeiras

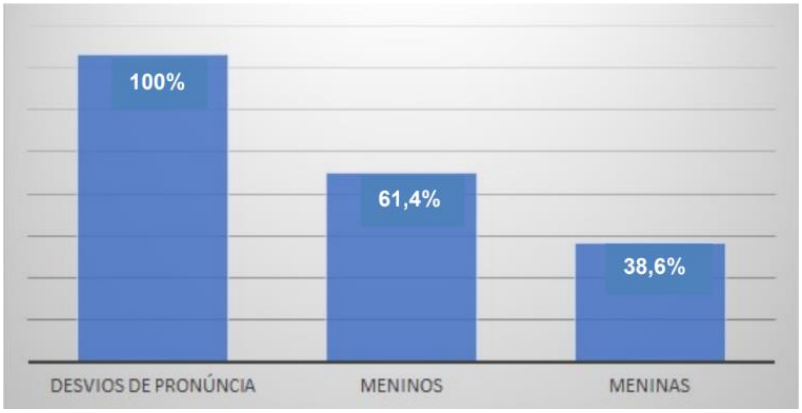
séries do Ensino Fundamental 1 (AREND, 2021) e a alfabetização em leitura por alunos pré-escolares (NEVES, 2021). Andrade (2022) utilizou a estrutura da base para desenvolver e implementar testes de recepção/percepção auditiva e compreensão oral.

Para o trabalho de Arend (2021), que utilizou informações do banco relativas a testes de produção oral, gerou-se uma *view* com dados específicos a ser manipulados e analisados.

Um dos resultados do trabalho de Arend (2021) é retratado na Figura 3, a seguir, em que os percentuais se referem ao total de desvios de pronúncia (3660) encontrados em crianças de 4, 5 e 6 anos, matriculadas no pré-escolar 1, pré-escolar 2 e primeiro ano do Ensino Fundamental, do qual os meninos tiveram 2247 desvios e as meninas, 1413.

A literatura elucidou os motivos dessa discrepância. Em Ramos (2008), o autor encontrou que meninos são mais afetados por desordens de linguagem do que meninas, por conta de alguns fatores, como a produção de testosterona no hemisfério esquerdo, responsável pela linguagem, que causa inibição em seu desenvolvimento. Já meninas desenvolvem o lado esquerdo do cérebro antes dos meninos, o que gera mais facilidade na aquisição da linguagem e alfabetização. Embora esse fato fosse conhecido, os percentuais revelados são inéditos, tendo em vista a quantidade de crianças estudadas (899). As ferramentas da Ciência de Dados ajudam a aprofundar a análise desses dados, especialmente com a extração de padrões.

Figura 3 – Desvios fonológicos em crianças de 4, 5 e 6 anos



Fonte: Arend (2021).

Os resultados trazidos nesta seção ainda podem ser complementados e ampliados, ou seja, a base de dados ainda está em estado praticamente bruto, pronta para ser objeto da ciência de dados, com suas ferramentas específicas e poderosas para análise de dados.

3 METODOLOGIA

Esta seção detalha: a base de dados que será utilizada na pesquisa; apresenta algumas das ferramentas computacionais de mineração de dados e aprendizagem de máquina que auxiliarão na análise dos dados da base; e também traz considerações metodológicas específicas.

3.1 TIPO DE PESQUISA

Tendo em vista as características da base de dados e as técnicas que são aplicadas, este trabalho é uma pesquisa que transita entre exploratória, descritiva e explicativa, que aqui se complementam.

De acordo com Gil (2002), a pesquisa exploratória busca, por meio de seus métodos e critérios, uma proximidade da realidade do objeto estudado. Proporciona maior familiaridade com o tema. É considerada uma metodologia aplicada inicialmente numa pesquisa científica, e posteriormente poderá ser melhor desenvolvida com a pesquisa descritiva. Entende-se que esta pesquisa será parcialmente exploratória, tendo em vista o conhecimento da pesquisadora sobre o tema e o conteúdo da base de dados, no entanto, há poucos estudos que abordem os temas da base.

O mesmo autor diz que a pesquisa descritiva realiza um estudo mais detalhado, com levantamento, análise e interpretação de dados. O pesquisador deve trabalhar como observador, mantendo-se distante do objeto de estudo, para que não influencie os resultados. Nesse modelo, as respostas resumem-se a dados principalmente quantitativos. A principal diferença entre a pesquisa descritiva e a exploratória está no conhecimento que o investigador tem sobre o objeto de estudo. Para manter-se diante do estudo, neste caso, deve-se olhar os dados a partir dos resultados das técnicas de ciência de dados aplicadas.

Finalmente, Gil (2002) diz que a pesquisa explicativa, além de realizar esse estudo aprofundado, também relaciona teoria e prática no processo da pesquisa. Identifica os fatores que determinam a ocorrência do fenômeno estudados ou que

contribuem para ele ocorrer. Trata-se de um modelo mais detalhado de estudo, assim, é considerada uma etapa avançada das pesquisas exploratórias e descritivas.

3.2 PROCESSAMENTO DA LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN ou NLP – Natural Language Processing) mescla Ciência da Computação, Inteligência Artificial e Linguística, dedicando-se à geração e compreensão automática da linguagem natural. Também, PLN é um campo de *machine learning* que lida com a capacidade de um computador ou de uma máquina de entender e analisar a linguagem humana. Por linguagem humana entende-se a fala, a escrita e os sinais. PLN é aplicado a diferentes áreas e tecnologias e cada área faz seu uso diferenciado de PLN (MANNING; SCHUTZE, 1999).

Dentre os componentes que formam a base de uma linguagem natural, estão a fonética e a fonologia – a segunda está vastamente presente nesta pesquisa. Em suma, a fonética e a fonologia tratam do som e das suas propriedades acústicas, como visto. Assim, este estudo utiliza princípios gerais de PLN, pois os dados de trabalho são linguísticos, traduzidos em caracteres fonológicos. No entanto, como não se abordam textos, muitas vezes, devido ao KDD, algoritmos mais genéricos podem responder melhor aos dados do que algoritmos específicos para PLN.

Na mineração de dados, há algoritmos como o Bag of Words, que permite gerar a nuvem de palavras (*word cloud*), por exemplo, que, apesar de simples, são funcionais, pois permitem uma visão dimensional da quantidade de ocorrências iguais presentes em um texto ou até *dataset*.

3.3 A BASE DE DADOS E O SISTEMA NHENHÉM FONOAUD (NhF)

A base de dados utilizada para desenvolvimento deste estudo possui mais de 30 mil registros, constituídos por mais de 30 campos. Os dados foram coletados desde março de 2019 até o momento, o que significa que a base cresce constantemente.

Os dados referem-se a testes realizados com crianças de 3 a 7 anos, a pedido da escola, da clínica fonoaudiológica ou de seus pais ou responsáveis, aplicados

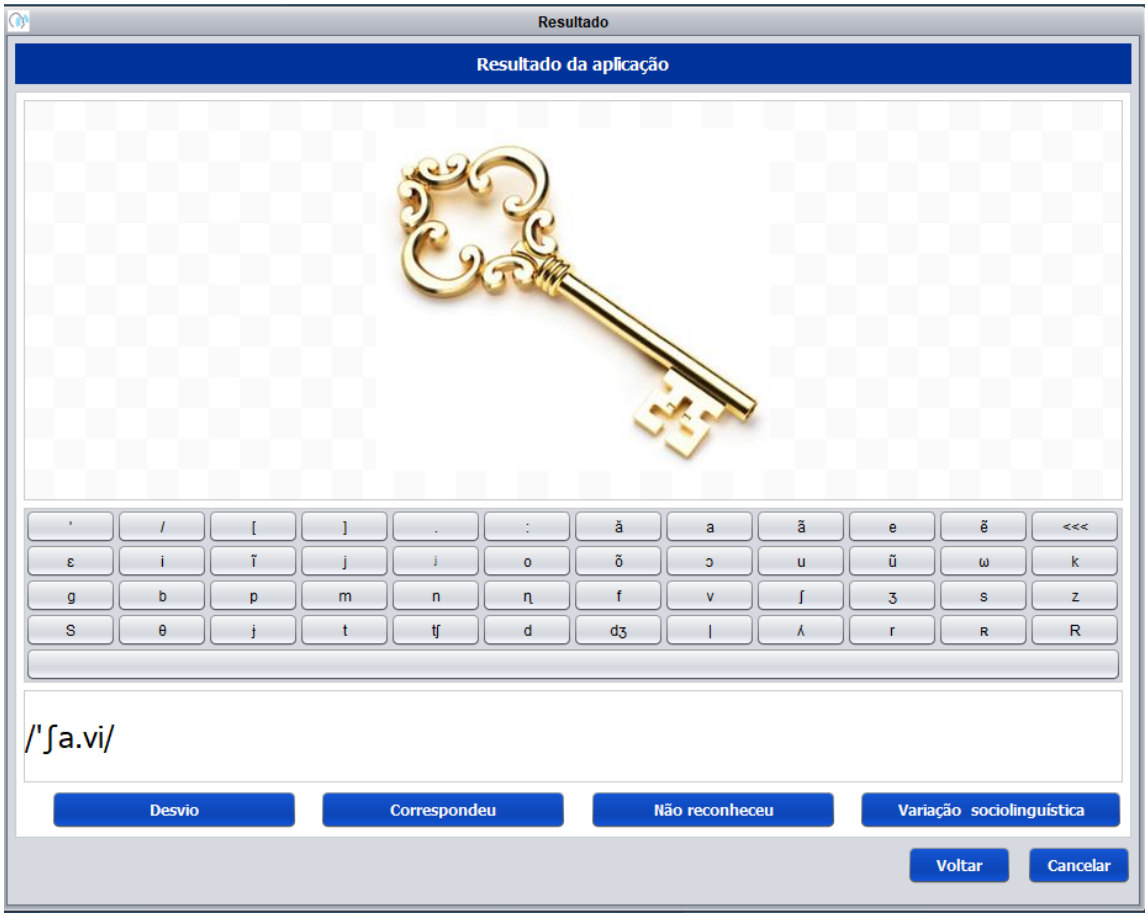
apenas por profissionais, com termo de autorização para utilização em pesquisa e garantia de anonimato, respeitando-se a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018). Seu formato é relacional, que é adequado para a mineração de dados. Assim, há certeza da procedência e forma de coleta dos dados, de modo que o banco está praticamente livre de inconsistências, como entradas ausentes ou incorretas, ou elas são bem conhecidas e podem ser propriamente trabalhadas.

A alimentação da base – ou seja, a coleta de dados – é feita por meio do *software desktop* Nhenhém Fonoaud (2014-2022). Utilizando o algoritmo fonológico-prosódico do conversor grafema-fonema Nhenhém (VASILÉVSKI, 2008), construiu-se o Nhenhém Fonoaud (NhF), um programa de computador para auxiliar na terapia fonoaudiológica, na aquisição da linguagem e na aprendizagem da leitura, o qual possibilita estudos longitudinais (VASILEVSKI; ARAÚJO; BLASI, 2014). Os dados são registrados em um formato fonológico-fonético, essencial para estudos em fala, uma vez que abordam os processos fonológicos. O NhF foi complementado ao longo dos últimos anos, para apresentar leiaute intuitivo e agradável ao usuário, suportar testes de percepção oral e leitura, bem como permitir pesquisa e geração de relatórios (FERNANDES; VASILEVSKI; ARAÚJO, 2020).

A Figura 4 mostra a tela de avaliação do NhF, em que se pode editar a transcrição fonológica de acordo com o que a criança falar e registrar o resultado de sua pronúncia.

O Nhenhém Fonoaud é um sistema desenvolvido em Java, conectado a uma base de dados PostGreSQL, uma ferramenta que atua como sistema de gerenciamento de bancos de dados relacionais. Seu foco é permitir implementação da linguagem SQL em estruturas, garantindo um trabalho com os padrões desse tipo de ordenação dos dados (SOUZA, 2020). O aplicativo de *software* cliente SQL e ferramenta de administração de banco de dados DBeaver facilita o acesso à base de dados NhF. O DBeaver é um programa multiplataforma, que tem por objetivo conectar e manipular vários tipos de banco de dados.

Figura 4 – Tela de registro do teste de produção oral do NhF, versão 2020

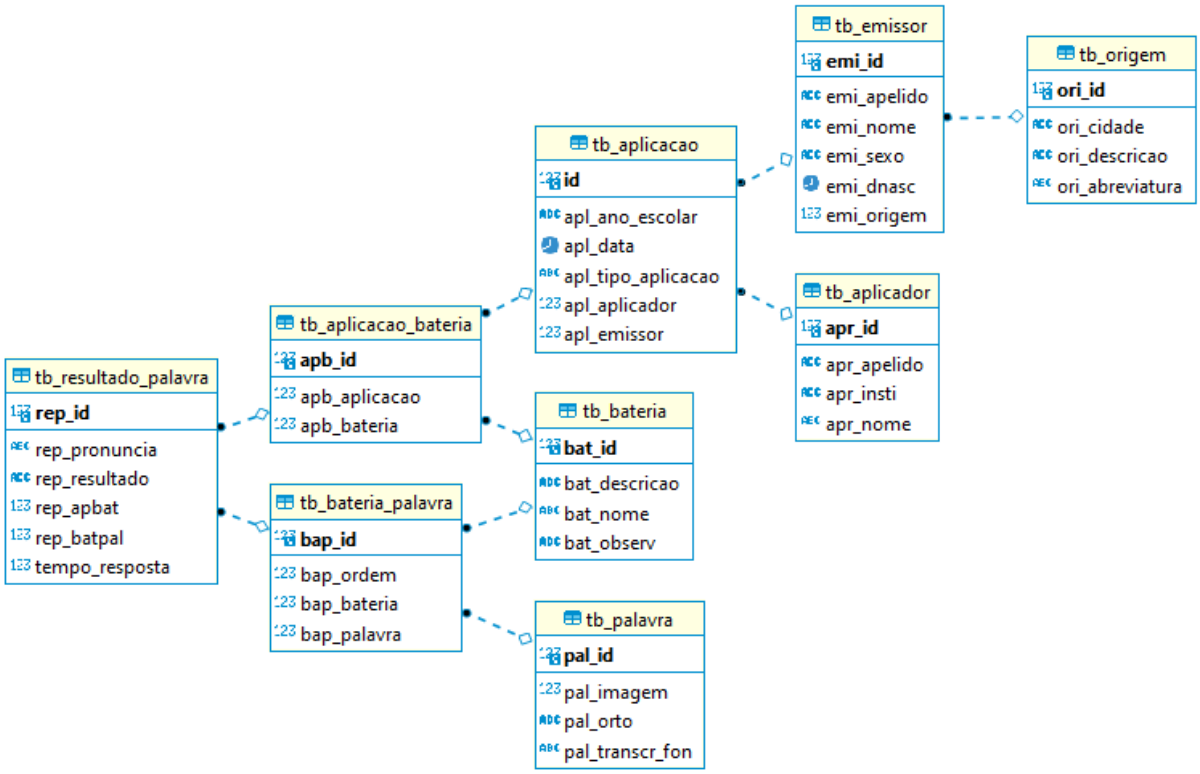


Fonte: Autoria própria.

Permite acessar bancos MySQL, PostGreSQL, Firebird, SQL Server, dentre outros. Em bancos de dados relacionais, o DBeaver usa a interface de programação de aplicativos JDBC para interagir com bancos de dados por meio de um *driver* JDBC (KINGHOST, 2022). Diretamente do DBeaver, foram extraídas as entidades (Figura 5), o fragmento de lista (Figura 6) e a *view* (Figura 2).

A organização da base de dados é a seguinte:

Figura 5 – Entidades originais do banco de dados



Fonte: Autoria própria.

Cabe destacar que é possível fazer muitas derivações a partir dos dados, criando-se vários arquivos específicos para tratar determinados temas. Afinal, a limpeza de dados envolve também a localização de correlações ocultas nos dados, a identificação de fontes de dados mais precisas e a determinação de quais colunas são mais apropriadas para as análises que se pretende realizar, pois trabalhar com um conjunto de dados muito grande pode dificultar o exame de todos os fatos relacionados a um problema que se quer estudar. Tal exame é importante para obter qualidade de dados. A base está pronta para a criação de perfis de dados e pode ser facilmente preparada para submissão a ferramentas automatizadas de limpeza e filtragem, dentre outras.

As informações provêm de testes de produção oral, que configuram o maior número de dados até o momento; seguidas do teste de leitura originado de aprendizagem via método fônico; e logo começará a ser alimentada com dados de percepção auditiva e compreensão de frases simples. A Figura 6 mostra um fragmento de registros do teste de produção oral de uma criança no banco de dados, em que se

destacam dados no formato grafofonêmico (transcrição fonológica), em que o campo ‘fonologia’ traz a pronúncia esperada para a ‘ortografia’, e o campo ‘pronuncia’ refere-se ao que o emissor, no caso, falou.

Cabe esclarecer que a base pesquisada contém transcrição fonológica, e não fonética, porque o objetivo é registrar os fonemas básicos do português, ou seja, as características essenciais de cada som (conforme Quadro 1 e Quadro 2), que realmente distinguem palavras, e não traços dialetais, portanto, não distintivos.

Figura 6 – Fragmento exibindo parte do resultado de um teste de produção oral

nascimento	sexo	ortografia	fonologia	pronuncia	resultado	tempo_resposta
2016-02-01	M	barata	/ba.'ra.ta/	/ba.'la.ta/	NAO_CORRESPONDEU	4,466
2016-02-01	M	vaca	/'va.ka/	/'va.ka/	CORRESPONDEU	2,825
2016-02-01	M	faca	/'fa.ka/	/'fa.ka/	CORRESPONDEU	1,984
2016-02-01	M	concha	/'kõ.ʃa/	/'kõ.sa/	NAO_CORRESPONDEU	12,637
2016-02-01	M	trança	/'trã.sa/	/'tã.sa/	NAO_CORRESPONDEU	5,809
2016-02-01	M	palhaço	/pa.'ʎa.su/	/paj.'a.su/	NAO_CORRESPONDEU	4,280
2016-02-01	M	cola	/'kɔ.la/	/'kɔ.ga/	NAO_CORRESPONDEU	10,177
2016-02-01	M	rosa (cor)	/'ʁɔ.za/	/'ʁɔ.za/	CORRESPONDEU	3,090
2016-02-01	M	palito	/pa.'li.tu/	/pa.'li.tu/	CORRESPONDEU	2,400
2016-02-01	M	unha	/'ũ.ɲa/	/'ũ.ɲa/	CORRESPONDEU	15,328
2016-02-01	M	dinheiro	/dĩ.'ɲej.ru/	/zĩ.'ɲej.lu/	NAO_CORRESPONDEU	17,000
2016-02-01	M	nuvem	/'nu.vẽj/	/'u.vẽj/	NAO_CORRESPONDEU	7,611
2016-02-01	M	massinha	/ma.'sĩ.ɲa/	/ma.'sĩ.ɲa/	CORRESPONDEU	2,019
2016-02-01	M	boca	/'bo.ka/	/'bo.ka/	CORRESPONDEU	8,786
2016-02-01	M	chave	/'ʃa.vi/	/'sa.vi/	NAO_CORRESPONDEU	7,272
2016-02-01	M	gelo	/'ʒe.lu/	/'ze.gu/	NAO_CORRESPONDEU	11,786
2016-02-01	M	zebra	/'ze.bra/	/'ze.ba/	NAO_CORRESPONDEU	6,903
2016-02-01	M	médico	/'mɛ.di.ku/	/'mɛ.di.ku/	CORRESPONDEU	2,812
2016-02-01	M	maçã	/ma.'sã/	/ma.'sã/	CORRESPONDEU	4,704
2016-02-01	M	pato	/'pa.tu/	/'pa.tu/	CORRESPONDEU	1,652
2016-02-01	M	flor (cluster)	/flor/	/foj/	NAO_CORRESPONDEU	5,520
2016-02-01	M	flor (coda)	/flor/	/foj/	NAO_CORRESPONDEU	5,000

Fonte: Autoria própria.

É válido expor que os caracteres fonético-fonológicos podem se desconfigurar no intercâmbio dos dados entre as ferramentas computacionais, da mesma forma que ocorre com caracteres acentuados, que podem não ser reconhecidos por alguns programas. Persistem nos dias atuais tais casos de desconfiguração, mas muito menos e mais facilmente tratáveis do que em épocas não tão remotas.

3.4 MODELAGEM DO DATA WAREHOUSE DE PRODUÇÃO ORAL

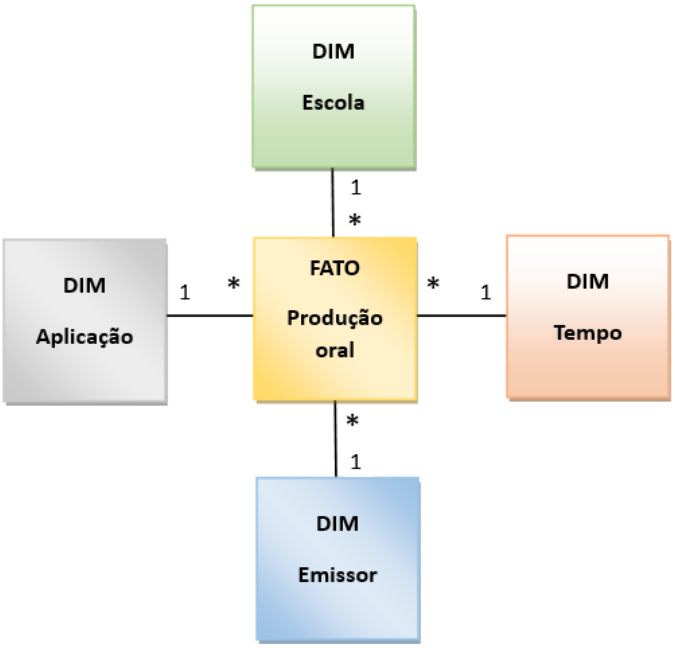
A partir da base NhF gerou-se um *data warehouse* apenas com a parte de produção oral, para nortear a análise desses dados, com a criação de campos específicos e ajustes necessários, sem alterar a base original, que é alimentada constantemente. Uma das fases mais importantes de um trabalho desta natureza é criar e preparar uma base de dados para pesquisar linguagem, que seja representativa da oralidade infantil e isso se garantiu na montagem da base NhF. Ainda, para ser eficiente, ela precisa ser compatível com ferramentas de *business intelligence* e mineração de dados, para que possa ser exportada e lida por essas ferramentas.

A parte dos dados utilizada na criação do *data warehouse* contém 26.578 registros de 1.098 crianças. Registra o ano escolar da criança, que pode ser Maternal 2 (MAT2), Pré-escolar 1 (PRE1), Pré-escolar 2 (PRE2) e 1.º ano do Ensino Fundamental (EF1). Na região pesquisada, onde o teste foi aplicado e os dados autorizados para pesquisa, tendo em vista a descendência alemã e italiana, é comum a não-realização do fonema /ʀ/, como em “ca**rr**o”, sendo substituído por /r/, como em “ca**r**o”. Essa variação provoca a neutralização entre esses dois fonemas, interferindo na alfabetização, uma vez que unifica pronúncias que são representadas por regras distintas na escrita.

Em termos de tabelas dimensão, fato e seus relacionamentos, o *dataset* está organizado conforme a Figura 7, na qual Aplicação se refere à sessão em que foi feito o teste, Emissor é a criança que fez o teste, Data é a data em que foi realizado o teste e a tabela fato contém os resultados dos testes de Produção Oral aplicados. Nesse sentido, destaca-se a modelagem multimensional, uma técnica de estrutura de dados otimizada para armazenamento em *data warehouse*. Seu objetivo é a otimização, para recuperação rápida e segura dos dados. Um modelo dimensional é projetado para ler, resumir, analisar informações numéricas. Os dados do *data warehouse* podem ser consultados via ferramentas Online Analytical Processing (OLAP), que são técnicas distintas para consulta analítica (MACHADO, 2013).

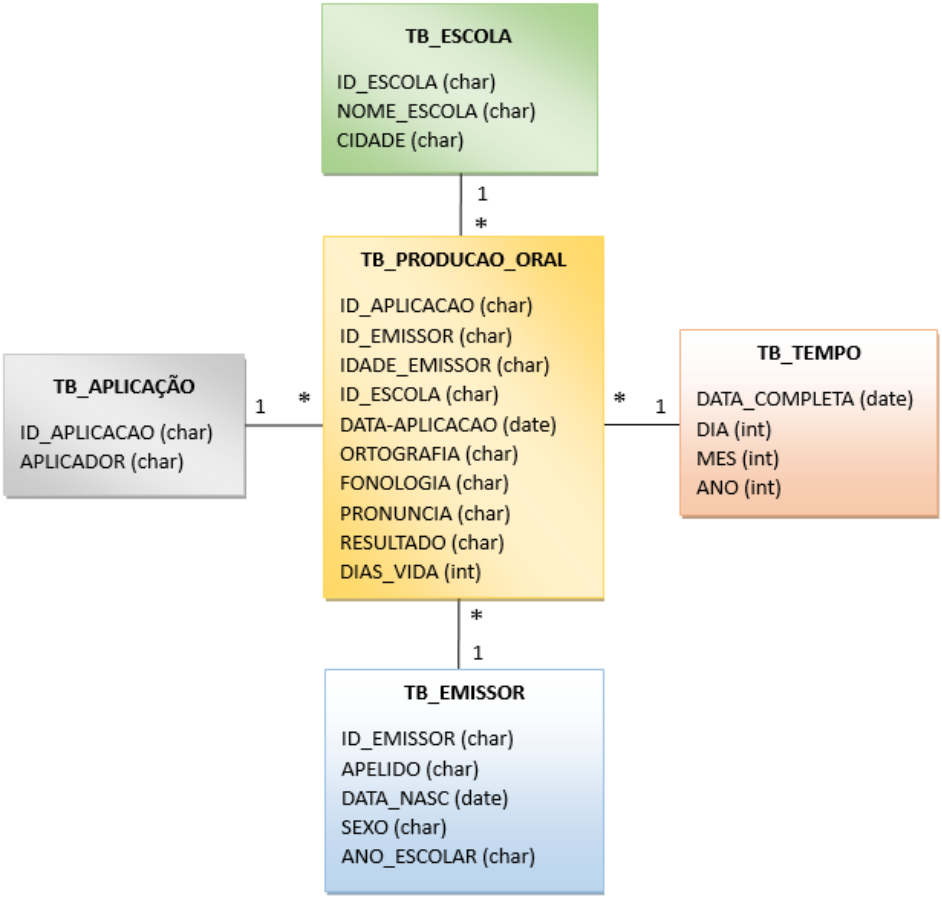
As entidades do banco de dados são compostas também pelos seguintes campos e respectivos tipos expressos na Figura 8.

Figura 7 – Modelo lógico



Fonte: Autoria própria.

Figura 8 – Modelo físico



Fonte: Autoria própria.

3.5 ETL

O processo de ETL ocorreu na etapa da divisão do banco original, em que se formou uma base nova, para ser editada sem alterar o banco de dados original e para conexão com a ferramenta de BI, o Power BI. Foram criados novos campos e as entidades continham dados específicos para trabalho com produção oral. Já nesse processo, algumas inconsistências foram encontradas e sanadas.

No Power BI, durante a criação de relatórios, outras inconsistências foram constatadas e ajustadas na própria base. À medida que se criavam gráficos, os resultados mostravam inconsistências, como em datas e alguns dados incompatíveis com o esperado. Sempre que isso ocorria, checagem e ajustes eram feitos no banco de dados e então atualizados para o PBI.

3.6 FERRAMENTAS DE MINERAÇÃO E ANÁLISE DE DADOS

A análise dos dados foi feita com o Power BI e o Orange, basicamente, por se mostrarem eficientes à proposta deste trabalho.

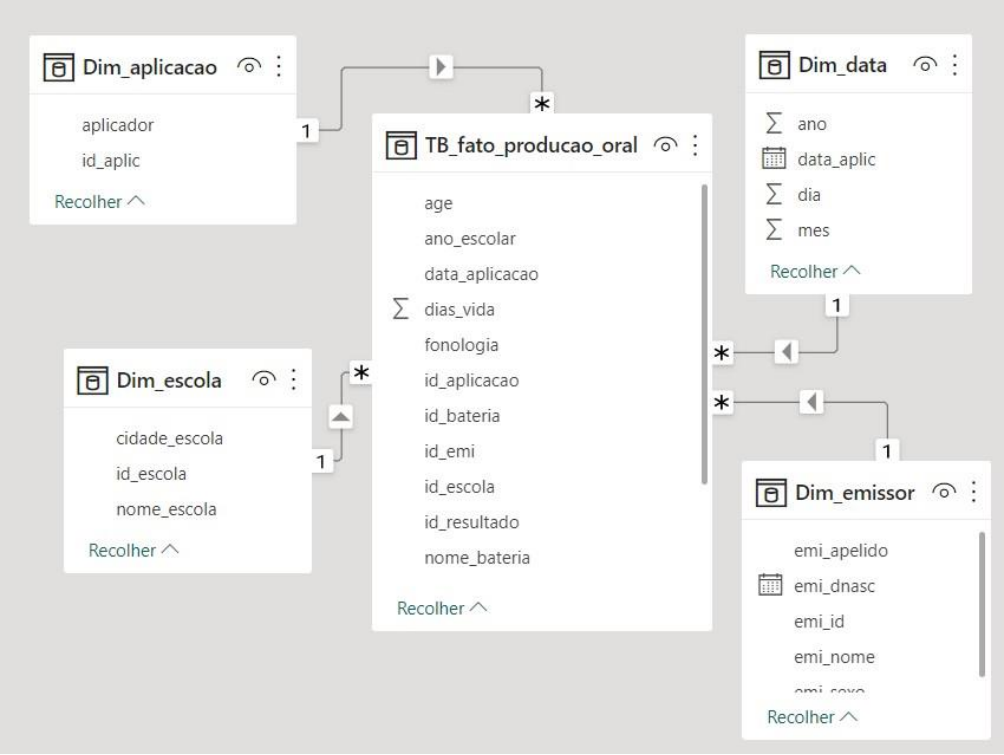
3.6.1 Power BI

Power BI (PBI) é uma ferramenta poderosa para análise de dados e geração de relatórios. Ela permite que as organizações transformem seus dados em informações significativas e tomem decisões baseadas em dados de maneira eficaz. Trata-se de uma plataforma unificada e escalonável para Business Intelligence empresarial e de autoatendimento. Conecta-se facilmente a dados de várias fontes e permite modelá-los e visualizá-los, criar relatórios de fácil personalização, inclusive com respostas rápidas impulsionadas por IA para perguntas sobre o problema em estudo (MICROSOFT, 2023b).

No PBI, os dados do *data warehouse* de produção oral foram carregados sem maiores problemas, e pouca transformação foi necessária. Os dados foram ajustados, quanto a alguns tipos, mas foi uma etapa rápida, uma vez que cuidados nesse sentido

foram tomados na modelagem da base no PostgreSQL. Os relacionamentos foram automaticamente reconhecidos pelo PBI (Figura 9).

Figura 9 – Exibição do Modelo no PBI



Fonte: Autoria própria.

3.6.2 Orange

Orange (2022) é um conjunto de ferramentas de visualização de dados de código aberto, aprendizado de máquina e mineração de dados. Ele apresenta um *front-end* de programação visual para análise exploratória rápida de dados qualitativos e visualização interativa de dados.

Orange consiste em uma interface de tela na qual o usuário coloca *widgets* e cria um fluxo de trabalho de análise de dados. Os *widgets* oferecem funcionalidades básicas como ler os dados, mostrar uma tabela de dados, selecionar recursos, treinar preditores, comparar algoritmos de aprendizado, visualizar elementos de dados etc. O usuário pode explorar visualizações interativamente ou alimentar o subconjunto selecionado em outros *widgets* (ORANGE, 2022).

Tais ferramentas possibilitam processar os dados, utilizando os recursos mencionados no Referencial Teórico. Os dados, após os procedimentos descritos na seção 2.7.1, 2.7.2, 3.4 e 3.5 foram convertidos e submetidos aos algoritmos de mineração de dados dessa ferramenta.

3.6.3 Algoritmos de Mineração de Dados

Vários algoritmos de *data mining* foram testados e mostraram sua validade. Cabe mencionar os três que apresentaram resultados mais pertinentes.

3.6.3.1 K-Means

K-Means é um algoritmo de clusterização (ou agrupamento), de aprendizado não supervisionado (que não precisa de *inputs* de confirmação externos), que avalia e agrupa os dados de acordo com suas características. É um método bastante utilizado (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

O k-Means ou k-médias define um protótipo em termos de um centroide, que é geralmente a média de um grupo de pontos, e é geralmente aplicada a objetos em um espaço n-dimensional contínuo, mas vem se mostrando eficiente quando aplicado a dados discretos e categóricos. Escolhem-se os centroides iniciais, em que k é um parâmetro especificado pelo usuário – o número de grupos desejado. Cada ponto é atribuído ao centroide mais próximo e cada coleção de pontos atribuídos a um centroide é um grupo. O centroide de cada grupo é então atualizado, com base nos pontos atribuídos ao grupo. Repetem-se os passos de atribuição e atualização até que nenhum ponto mude de grupo ou, equivalentemente, até que os centroides permaneçam os mesmos (TAN; STEINBACH; KUMAR, 2009).

Nesse sentido, a análise por Silhouette define o quão bem um ponto se encaixa em um *cluster*, de modo que um gráfico mede o quão perto os pontos de um *cluster* estão dos pontos de outro *cluster* mais próximo. O coeficiente de Silhouette, quando próximo de +1, indica que os pontos estão muito longe dos pontos do outro *cluster* e, quando próximo de 0, indica que os pontos então muito perto ou até interseccionando

outro *cluster* (HAN; KAMBER; PEI, 2011). A situação ideal, portanto, é obter pontos o mais próximo de 1 possível e mais distante de 0 possível.

3.6.3.2 Associação

A fim de encontrar relações entre ocorrências nos dados, usa-se análise de associação, também algoritmos não supervisionados. Devido a sua grande aplicabilidade, as regras de associação encontram-se entre os mais importantes tipos de conhecimento que podem ser minerados em bases de dados. Associação ou descoberta de regras associativas encontra subconjuntos de itens que ocorrem de forma simultânea e frequentemente em uma fração mínima e previamente estabelecida do conjunto de dados.

A força da regra de associação é medida pelo seu suporte e confiança. Esse processo leva em conta a confiança mínima, que se refere a um valor predeterminado que expressa a qualidade de uma regra, ou seja, o quanto a ocorrência do antecedente da regra pode assegurar a ocorrência do consequente da regra. Uma regra de associação é considerada válida se o número de vezes em que ela ocorre for superior à confiança mínima estabelecida (TAN; STEINBACH; KUMAR, 2009).

$$(1)\ X \rightarrow Y$$

$$(2)\ Y \wedge W \rightarrow Z$$

Nesse caso, a ocorrência de X induz à ocorrência de Y (1) e a ocorrência de Y e W induz à ocorrência de Z. A medida de suporte representa a porcentagem de registros da base de dados que contêm os itens X e Y, indicando sua relevância. Já a confiança representa, dentre os registros que possuem o item X, a porcentagem de registros que possuem também o item Y, indicando a validade da regra. Nesse sentido, a meta da mineração de regras de associação é encontrar todas as regras que tenham valores de suporte e de confiança iguais ou superiores aos valores mínimos definidos pelo usuário.

3.6.3.3 Naïve Bayes

Diversos algoritmos de mineração de dados são fundamentais em princípios e teorias estatísticas. Um deles é o Naïve Bayes, classificador Bayesiano ingênuo (CBI), que se baseia no Teorema de Bayes, portanto, relaciona-se ao cálculo de probabilidades condicionais. Esse classificador multinomial é bastante utilizado no aprendizado de máquina. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes predefinidas. O termo *naïve* (ingênuo) refere-se à premissa central do algoritmo, cuja hipótese é de que os atributos considerados não se correlacionam entre si, ou seja, são independentes, o que, em muitos casos, não ocorre (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Um classificador necessita de um conjunto de treino e um de teste, o que pode ser feito dividindo-se o *dataset* de pesquisa. O procedimento mais comum é a validação cruzada, que divide os dados em k conjuntos e usa k - 1 conjuntos para treinamento e o conjunto restante para teste. Esse procedimento é repetido, de modo que cada conjunto tenha sido usado para teste exatamente uma vez. O resultado relatado é a precisão média do modelo (HAN; KAMBER; PEI, 2011). A validação cruzada é usada para detectar sobreajuste (*overfitting*), ou seja, a não-generalização de um padrão. Ocorre quando, no conjunto de treino, o modelo tem desempenho excelente, porém, no conjunto de teste, o resultado é ruim.

Para visualização do desempenho de um algoritmo de classificação utiliza-se uma tabela chamada matriz de confusão. Essa tabela, também chamada de matriz de erro, traz as previsões corretas e incorretas do modelo (Figura 10):

Figura 10 – Matriz de confusão 2x2

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Microsoft (2023a).

- Verdadeiros Positivos: classificação correta da classe Positivo;
- Falsos Negativos (Erro Tipo II): erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo;
- Falsos Positivos (Erro Tipo I): erro em que o modelo previu a classe Positivo quando o valor real era classe Negativo;
- Verdadeiros Negativos: classificação correta da classe Negativo.

Várias métricas são utilizadas para avaliar o modelo, especialmente acurácia, que indica a *performance* geral do modelo, ou seja, dentre todas as classificações, quantas ele classificou corretamente; precisão, que indica, dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas; *recall*, que indica, dentre todas as situações de classe Positivo com o valor esperado, quantas estão corretas; e F1 *score*, média harmônica entre precisão e *recall*.

3.7 INTERPRETAÇÃO

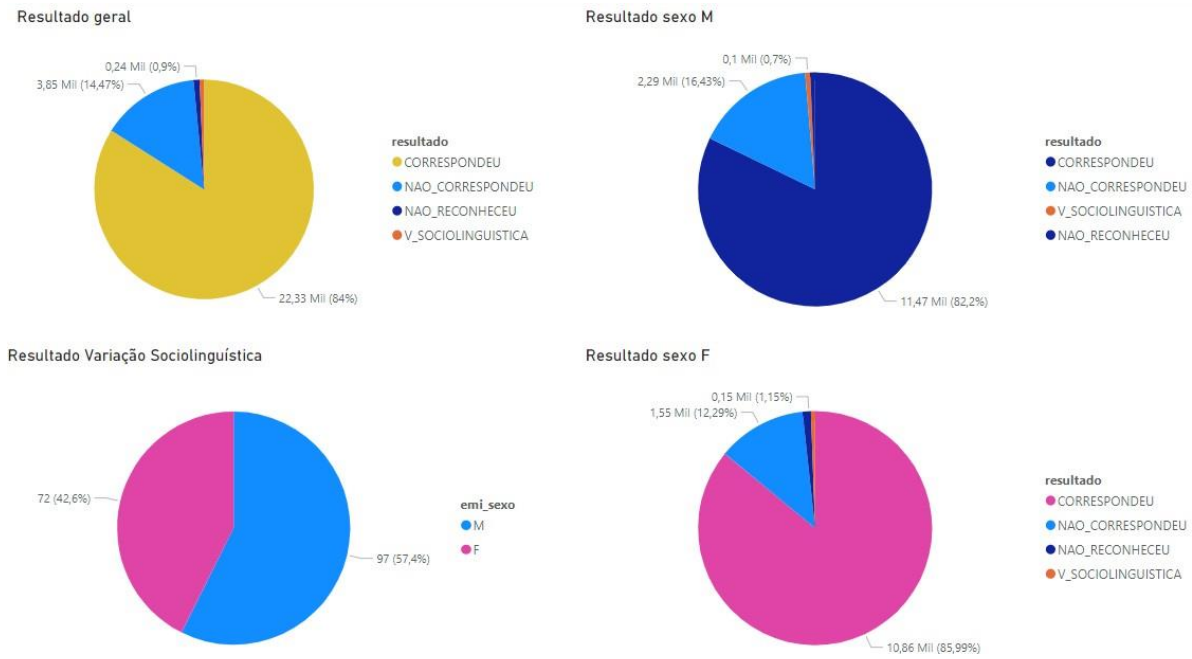
Finalmente, após todo esse processo, houve o pós-processamento em si: a interpretação das características, regras e dos padrões encontrados, que é decisiva para a relevância da pesquisa e da aplicação de seus resultados. Nesse ponto, destacaram-se conhecimentos específicos das Neurociências e da Linguística sobre aquisição da linguagem, principalmente.

Em suma, dessa interpretação decorrem a descoberta do conhecimento e sua validação e avalia-se a atuação da Ciência de Dados frente a dados linguísticos, sobretudo à parte fonológica.

4 RESULTADOS

Muitas foram as possibilidades geradas pelos relatórios no PBI. Algumas delas foram escolhidas para esta análise, como os resultados gerais de fonoarticulação, que são os seguintes.

Figura 11 – Totais de articulação gerais, por sexo e variação sociolinguística

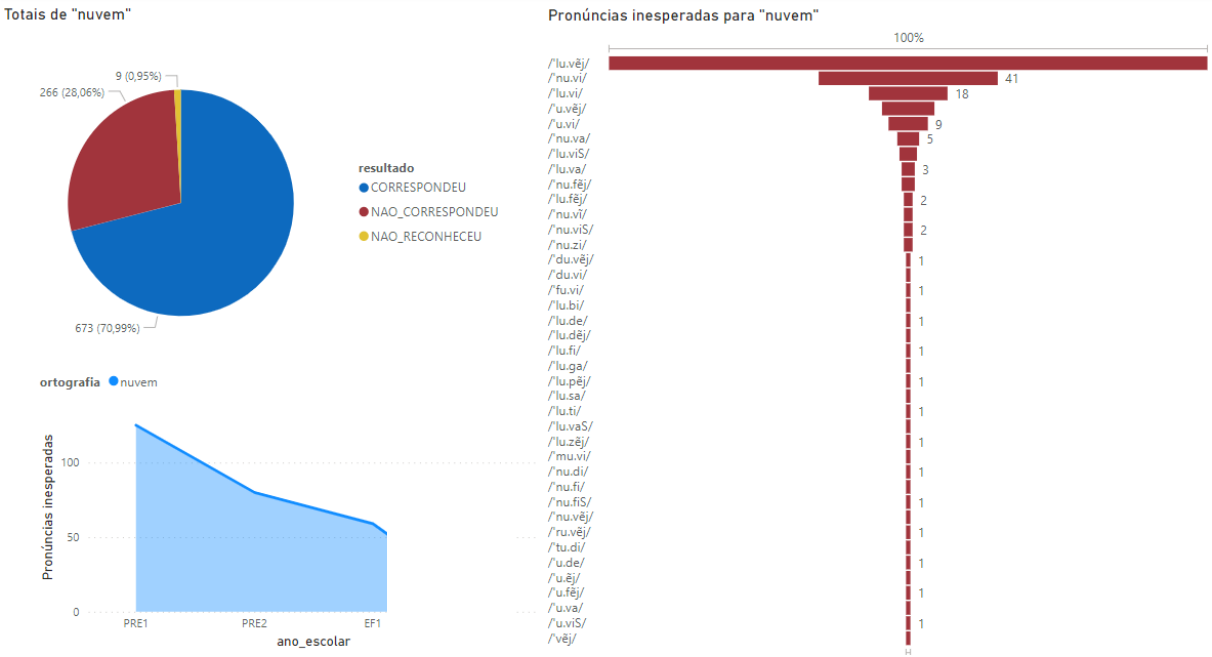


Fonte: Dados primários.

A partir da Figura 11, percebe-se que a grande maioria das crianças participantes já amadureceu totalmente a fonoarticulação, com leve vantagem para as meninas. A variação sociolinguística ocorre em menos de 1% dos dados coletados e os meninos são levemente mais propensos a ela. O percentual pequeno de NÃO_RECONHECEU referenda e valida a escolha das figuras (imagens) para o teste e a disposição das crianças a participar.

Durante a análise, vários fenômenos linguísticos notáveis ocorreram. As palavras “dinheiro” e “nuvem” foram as que mais mostraram pronúncias variadas. A palavra “nuvem” teve 39 pronúncias diferentes, dentre 266 pronúncias que não corresponderam ao esperado.

Figura 12 – Resultados para a pronúncia de “nuvem”



Fonte: Dados primários.

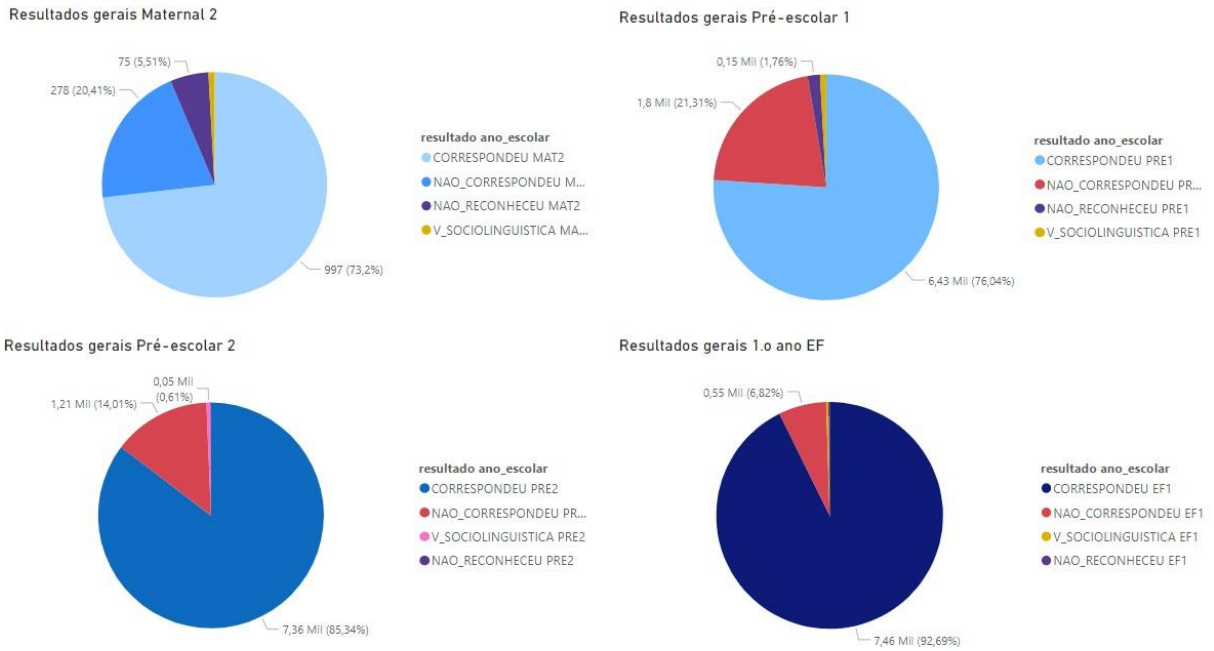
Na Figura 12, todas as pronúncias inesperadas tratam de processos fonológicos, uma vez que o objetivo era o emissor falar a palavra que ouviu como /'nu.vẽj/. A primeira variação que mostra a Figura 12, /'lu.vẽj/, ocorreu 137 vezes. Esse fato chama a atenção, porque não se constata a pronúncia /'lu.vẽj/ por adolescentes na cidade de onde os dados foram coletados, o que indica que se trata de uma variação relativa à idade, que deixará de ocorrer em pouco tempo, pois diminui conforme a escolarização aumenta. Sobre essa emissão, entende-se que ocorre um processo fonológico que pode ser chamado de lateralização de oclusiva nasal alveolar, que não está identificado no Quadro 3. Já /'nu.vi/, que ocorreu 41 vezes, traz a redução do ditongo nasal átono /ẽj/ para a vogal oral /i/. A próxima ocorreu 18 vezes, /'lu.vi/, e contém os dois processos fonológicos mencionados. A seguinte, /'u.vẽj/, ocorreu 12 vezes, e traz um caso de simplificação (redução) de consoante inicial. Então, com 9 ocorrências, há /'u.vi/, com simplificação de consoante inicial e ditongo final. Finalmente, observa-se que 13 crianças pronunciaram “nuvem” de maneira única, o que revela sua concepção particular do que ouvem e conseguem repetir.

Em se tratando de resultados gerais por ano escolar, que é de interesse maior neste estudo, obteve-se um relatório que mostra a quantidade de ocorrências e o percentual relativo a ela para cada fase (Figura 13).

Como há menos dados para o maternal 2 – fase que antecede o pré-escolar 1 –, os dados relativos a essa fase, apesar de reais e potencialmente reveladores, devem ser vistos com parcimônia.

A variação sociolinguística vai de 0,88% e 0,89% no maternal e pré-escolar 1, respectivamente, para 0,68% no pré-escolar 2 e 0,35% no primeiro ano do Ensino Fundamental. Da mesma forma, o número de correspondências (pronúncia igual à esperada) aumenta no decorrer dos anos escolares: em torno de 3% do maternal para o pré-escolar 1, 9% do pré-escolar 1 para o pré-escolar 2 e 7% do pré-escolar 2 para o primeiro ano do Ensino Fundamental. Apesar disso, em torno de 7% das crianças chegam ao EF1 (têm 6 ou 7 anos) sem ter adquirido plenamente o sistema fonológico da língua materna.

Figura 13 – Totais de fonoarticulação por ano escolar



Fonte: Dados primários.

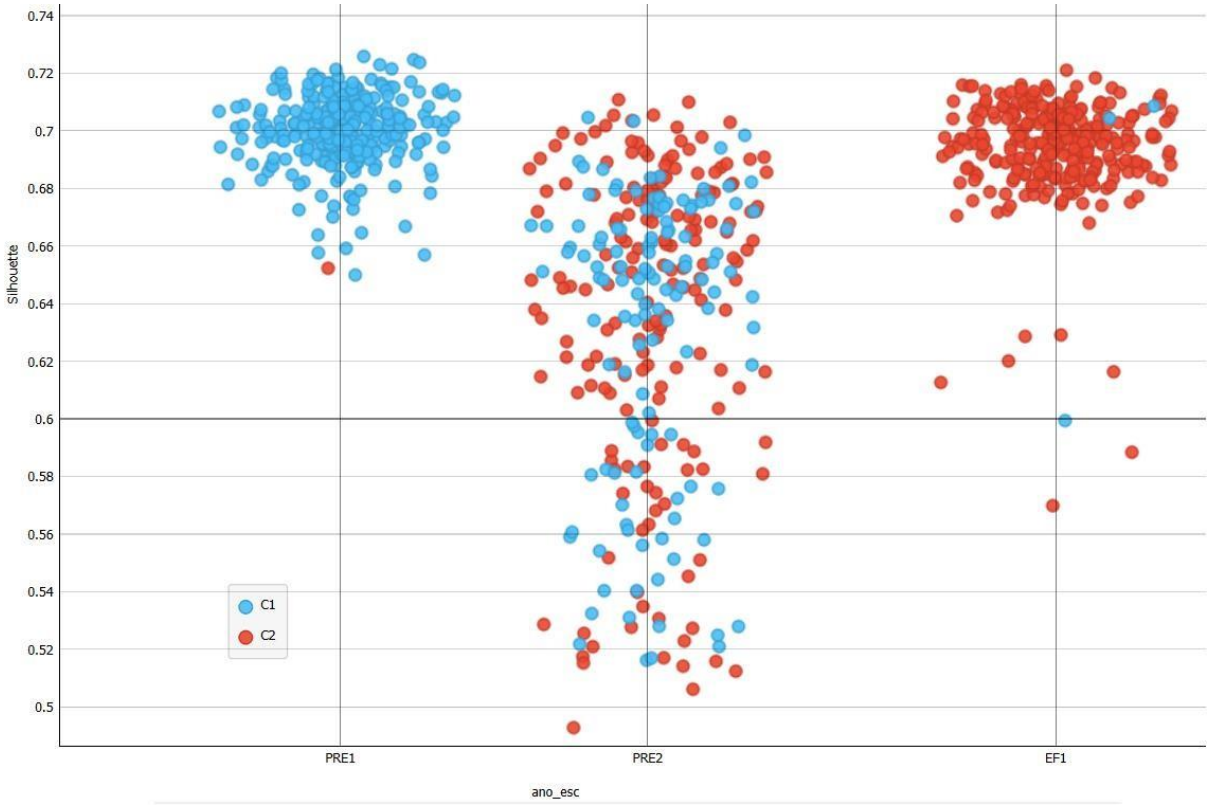
Ao processar dados no Orange, fez-se necessário criar campos derivados de outros campos e também campos calculados e condensar dados, conforme previsto. Utilizou-se a planilha eletrônica Excel para esse fim, cujos arquivos de dados são

perfeitamente carregados pelo Orange, para aplicação de algoritmos de mineração de dados. Os dados foram exportados para o Excel e então tratados.

Então, algumas questões relativas a características comuns aos registros foram tratadas com mineração de dados. Para tanto, um subconjunto do *dataset* foi selecionado, contendo dados de alunos do pré-escolar 1, pré-escolar 2 e 1.º ano do Ensino Fundamental apenas, que responderam à bateria de teste mais completa do *dataset*. Esses dados são maioria (23.946 registros) e foram então agrupados gerando novos atributos (campos calculados), totalizando 921 ocorrências. Calcularam-se o total de fonemas que a criança já adquiriu, a articulação dos encontros consonantais perfeitos e da coda silábica r interna e externa.

No Orange, o k-Means foi utilizado para verificar as condições de aquisição do sistema fonológico, quando a criança já está na escola. Levando em conta a quantidade de fonemas que falta adquirir, a idade da criança, a situação da aquisição e o ano escolar, o algoritmo encontrou 2 *clusters* com similaridade de 0,633 (Silhouette *scores*).

Figura 14 – Agrupamentos por situação da aquisição da linguagem oral nas séries escolares iniciais

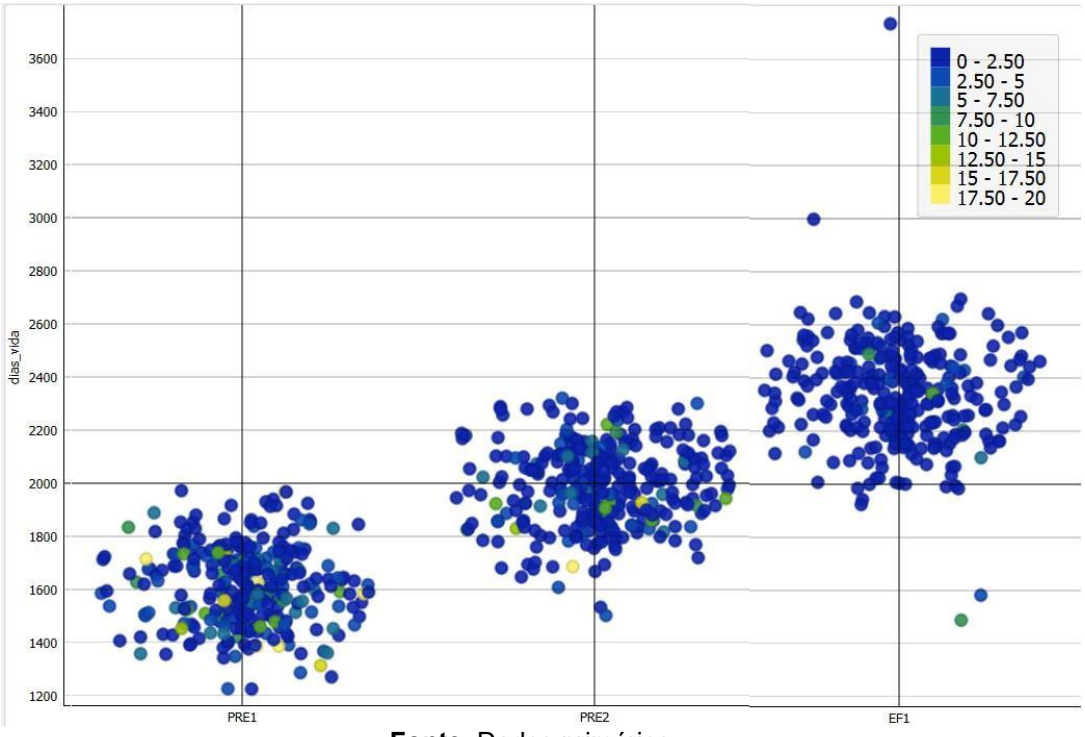


Fonte: Dados primários.

Considerando-se os *clusters* encontrados em relação ao ano escolar (Figura 14), compreende-se que o pré-escolar 2 – em torno de 5 anos de idade – é uma fase de transição na linguagem, heterogênea (vários casos discrepantes), em que muitas crianças não finalizaram a aquisição dos fonemas, muitas o finalizaram e entre esses extremos há vários outros níveis de aquisição. Por conta disso, há, nesse ano escolar, desde Silhouette próxima de 0 (não desejável) e próxima de 0,7 (desejável), enquanto no pré-escolar 1 e 1.º ano do EF a grande maioria dos dados está concentrada próximo de 0,7. Cada ponto do gráfico da Figura 14 refere-se a uma criança. Pontos discrepantes mostram casos específicos que devem ser analisados.

Ainda, em termos de fonemas que falta adquirir, o k-Means mostrou que, ao entrar na escola, a maioria das crianças já produz quase todos os sons de sua língua, apesar de haver crianças no EF1 para as quais em 2 anos de escola seus órgãos fonoarticuladores não amadureceram plenamente (Figura 15). Fica evidente que, no PRE1, para uma grande quantidade de crianças, falta adquirir mais de 7 fonemas ainda e, no EF1, persistem alguns casos desses, que, então, precisam de atenção por parte da escola.

Figura 15 – Quantidade de fonemas ainda não adquiridos



Fonte: Dados primários.

O Apriori, um algoritmo de associação, também foi aplicado por meio do Orange, a fim de verificar se havia alguma relação entre as pronúncias mais complexas da língua, ou seja, os encontros consonantais perfeitos (*clusters* fonológicos) e a coda silábica r, e a pronúncia de todos os fonemas. Isso é relevante porque há situações em que a criança pronuncia todos os fonemas, mas não consegue articular ambos esses casos complexos ou um deles.

Foram encontradas algumas relações relevantes, sobretudo no caso das codas silábicas r interna e final (Figura 16).

Figura 16 – Associação entre articulação da coda silábica r interna e externa e encontro consonantal perfeito

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.732	0.951	0.770	1.011	1.221	0.133	pron_codar_fim=S	→	pron_codar_meio=S
0.732	0.940	0.779	0.989	1.221	0.133	pron_codar_meio=S	→	pron_codar_fim=S
0.633	0.813	0.779	0.859	1.216	0.112	pron_codar_meio=S	→	pron_cluster_r=S
0.633	0.946	0.669	1.164	1.216	0.112	pron_cluster_r=S	→	pron_codar_meio=S
0.620	0.805	0.770	0.869	1.204	0.105	pron_codar_fim=S	→	pron_cluster_r=S
0.620	0.927	0.669	1.151	1.204	0.105	pron_cluster_r=S	→	pron_codar_fim=S
0.610	0.834	0.732	0.914	1.247	0.121	pron_codar_meio=S, pron_codar_fim=S	→	pron_cluster_r=S
0.610	0.984	0.620	1.256	1.264	0.128	pron_cluster_r=S, pron_codar_fim=S	→	pron_codar_meio=S
0.610	0.964	0.633	1.216	1.252	0.123	pron_cluster_r=S, pron_codar_meio=S	→	pron_codar_fim=S
0.610	0.912	0.669	1.094	1.247	0.121	pron_cluster_r=S	→	pron_codar_meio=S, pron_codar_fim=S

Fonte: Dados primários.

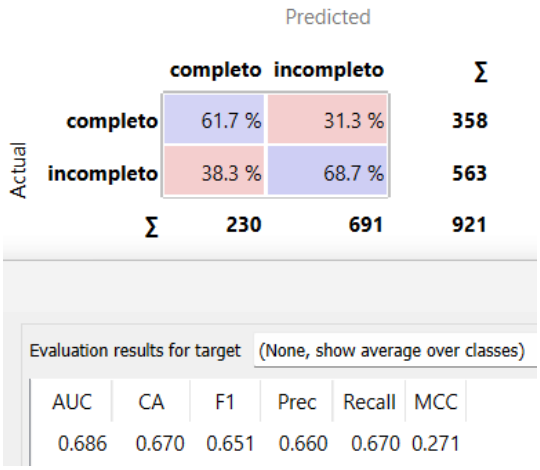
Constata-se, com suporte de 73% e confiança de 95%, que existe relação entre a articulação da coda silábica interna e coda silábica no fim da palavra. Dito de outro modo, obteve-se uma regra de associação que registra a maturidade articulatória para produzir a coda silábica r interna e em fim de palavra na base, dada por: {produz coda silábica r fim} Þ {produz coda silábica r meio}. Essa regra de associação indica que as crianças que articulam a {coda silábica r final} têm maior chance de também articular a {coda silábica r interna}.

Também existe associação entre a pronúncia das duas codas e do encontro consonantal com r, com suporte de 60%. Ainda, o encontro consonantal com l não aparece em associação com o nível de suporte mínimo de 60%. Já sobre o encontro com r, verifica-se que a criança que pronuncia a coda r no meio da palavra tem mais chance de também pronunciá-lo. Outras relações que aparecem na Figura 16 são até

parecidas, no entanto, prefere-se considerar os suportes maiores para aumentar a validade da regra. No *dataset*, as palavras que representavam as codas e encontro consonantal com r eram “po**r**ta”, “flo**r**” e “**tr**ança”.

Por fim, aplicou-se o Naïve Bayes a essa mesma base de 921 registros, a fim de verificar quais variáveis mais interferem na aquisição da linguagem oral. O alvo era identificar características que interferem na aquisição final (sistema fonológico completo) da linguagem oral. Testaram-se algumas combinações de variáveis e percebeu-se que o melhor desempenho é quando se considera a idade isoladamente (dias de vida) e idade junto com sexo (Figura 17):

Figura 17 – Matriz de confusão e métricas para o Naïve Bayes: situação (alvo) em relação a dias de vida



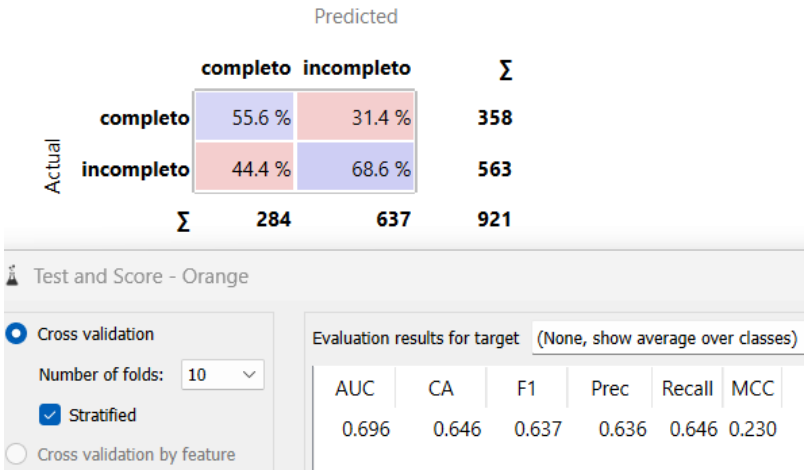
Fonte: Dados primários.

Observa-se que, apesar de o modelo representado na Figura 17 ter percentuais e métricas consideráveis, ele classifica apenas 230 ocorrências dentre 358, como situação “completo”. Dessas 230, 61,7% estão corretas.

Quando se considera, além dos dias de vida, o sexo do emissor, o Naïve Bayes classifica 284 ocorrências, dentre as 358, como situação “completo”, e delas 55,6% estão corretas (Figura 18). Nesse segundo caso, todas as métricas do modelo diminuem sensivelmente, contudo, o resultado do algoritmo reforça uma relação estreita entre a aquisição da linguagem verbal oral e as variáveis sexo e idade, que pode ser mais explorada e detalhada.

A questão da idade isoladamente é fato óbvio, pois depende do fator maturacional, contudo, nem isso foi satisfatoriamente medido ainda.

Figura 18 – Matriz de confusão e métricas para o Naïve Bayes: situação (alvo) em relação a dias de vida e sexo



Fonte: Dados primários.

5 DISCUSSÃO

Após esta jornada, cabe discutir alguns resultados – ou até mesmo descobertas – expressivos no tocante a aquisição da lingua verbal oral, mais precisamente, fonoarticulação; à construção do *data warehouse*; e à aplicação da mineração de dados ao *dataset* linguístico.

5.1 FONOARTICULAÇÃO

A partir de uma base de dados linguísticos de produção oral de crianças de 3 a 6-7 anos, este estudo mostra, com apoio de relatórios do Power BI e algoritmos de mineração de dados, características relevantes da aquisição do sistema fonológico. Os resultados indicam que a variação sociolinguística característica da região está diminuindo, pois a cada ano menos crianças a realizam, preferindo o fonema r forte. Aliás, os processos fonológicos encontrados nas variações de /'nu.vẽj/ levam a crer que se podem sugerir alterações nos processos fonológicos mais comuns na atualidade descritos no Quadro 3, tendo em vista que ele não traz um processo comum na região pesquisada.

A idade de 5 anos, em que a criança está no pré-escolar 2, parece complexa, havendo várias discrepâncias na aquisição dos fonemas, mas a maioria das crianças supera isso no ano seguinte. Mesmo assim, persistem casos de crianças que estão em idade de alfabetização e ainda não completaram a aquisição da linguagem verbal oral, de modo que a escola precisa conhecer essa realidade, a fim de prever e tratar dificuldades na aprendizagem da escrita. Também é válido saber se essas crianças distinguem tais fonemas, por meio de teste de audição simples.

Em relação à coda r em sílaba interna e em fim de palavra e aos encontros consonantais perfeitos, encontraram-se associações importantes, contudo, linguisticamente, chama atenção o fato de várias crianças conseguirem articular apenas um desses dois casos, o que levanta um tópico para investigação mais profunda, em busca de uma lógica (regra).

A relação estreita entre a completa aquisição do sistema fonológico e o sexo

(além da idade), sugerida pelo classificador Naïve Bayes, levanta a possibilidade de aprimorar o modelo, com mais dados e até outras variáveis, para adaptá-lo a analisar casos específicos com nível elevado de acerto. Enfim, constata-se que o classificador é adequado para abordar problemas linguísticos.

Dentre tantas possibilidades que a base de dados permite vislumbrar, como estudo futuro, sugere-se descobrir regras para os fonemas que são adquiridos mais tardiamente, a fim de conhecer mais as dificuldades de articulação, pois se observaram vários casos em que a criança pronuncia o r fraco, mas não pronuncia o fonema /ʎ/, como em “pa**lh**aço”, tendo mais de 4 anos.

O trabalho mostrou que é possível delinear padrões de aquisição da língua verbal oral, contudo, são necessários mais processamentos do que os que couberam neste estudo. Foram descobertas aqui características importantes, que sugerem prosseguir na busca por padrões, que pode ser concluída em estudos futuros, pois aqui se explora apenas pequena parte da base NhF.

5.2 DATA WAREHOUSE

A base de dados de produção oral modelada mostrou potencial de pesquisa muito maior do que foi explorado neste trabalho, de modo que o que aqui se expôs acaba sendo um início de investigação, cabendo explorá-la em vários trabalhos futuros, com várias temáticas relacionadas à aquisição da língua oral.

Ainda, como a base NhF aumenta, com a inserção de novos registros, advindos de novos testes, e tais dados são atualizados no PBI, podem-se acrescer registros dos mesmos testes feitos em crianças de idades diferentes das que há na base atualmente, bem como aumentar dados do Maternal 2, para fins de comparação e análise longitudinal. Isso permite mais extração de características das situações de que tratam os dados e, então, extração de padrões de aquisição do português brasileiro oral. Esse incremento pode melhorar o desempenho dos algoritmos de mineração de dados, a fim de obter modelos e características mais apurados.

Durante a manipulação de dados, constataram-se todos os processos mencionados no Quadro 3, inclusive os não esperados para o desenvolvimento.

Isso possibilita estudos que levem em conta sexo, idade e ano escolar da criança e o processo fonológico e muitos outros dados calculados. Seria possível, por exemplo, propor atualização no Quadro 3, a partir de uma nova abordagem tecnológica.

Apesar de parecer que os dados provenientes do PBI podiam ser criados em várias outras ferramentas, cabe lembrar que, como o banco foi exportado, ele atualiza automaticamente as alterações nos dados, inclusive valores em campos calculados, e as transfere aos gráficos gerados. Ainda, aos dados foram aplicados filtros diversos. Isso tudo não é possível em ferramentas não específicas para BI.

5.3 CIÊNCIA DE DADOS E DADOS FONOLÓGICOS

O caractere fonológico ʀ, que representa o fonema r forte, é um caractere r versalete. Então, foi intercambiado como r na importação do *data warehouse* pelo PBI, o que obscureceu o reconhecimento dos dados, no entanto, esse caso foi identificado e abordado propriamente na análise. Isso reforça a necessidade de conhecer a base de dados e seu conteúdo para trabalhar com as ferramentas de BI e mineração de dados.

Há outras barreiras para o processamento de dados fonológicos, tendo em vista que alguns algoritmos não conseguem ler os caracteres IPA. Por exemplo, a tentativa de usar mineração de texto com o *dataset* não foi satisfatória, pois não foi possível criar uma *bag of words* (*word cloud*) com os dados de variação sociolinguística de “nuvem” (Figura 19).

Apesar disso, de maneira geral, aplicação de Ciência de Dados ao *dataset* linguístico revela-se promissora, posto que, neste primeiro estudo, abordaram-se os dados de maneira diferenciada, cujos resultados apontam para pesquisas frutíferas na área da aquisição da linguagem verbal oral, pois as ferramentas aplicadas se revelaram adequadas para processar dados com transcrição fonológica, com poucos problemas no intercâmbio de caracteres do Alfabeto Fonético Internacional, os quais não prejudicam as análises, se o pesquisador conhece a base e a área alvo da pesquisa.

Figura 19 – *Word cloud* gerada para as pronúncias inesperadas de /'nu.vẽj/

Fonte: Dados primários.

Nessa linha, abre-se um norte para coletar e processar dados de testes de audição e de leitura, outras duas áreas da linguagem para as quais a base NhF está preparada.

CONSIDERAÇÕES FINAIS

Em seu propósito de auxiliar na compreensão das características que norteiam a aquisição da linguagem verbal oral, a partir da exploração de uma base de dados de fala, utilizando princípios, técnicas e ferramentas da Ciência de Dados, considera-se que este estudo foi bem-sucedido.

Resta comprovada a viabilidade de pesquisa relacionando Ciência de Dados e Linguística, sobretudo no que tange à fonologia, que utiliza caracteres específicos, nem sempre legíveis por *softwares*. *Data warehouse* e mineração de dados foram aplicados e resultados foram obtidos. Dificuldades foram documentadas. Sugestões de estudos futuros foram fornecidas. Evidenciou-se o grande potencial de investigação do *data warehouse* criado.

Assim, passos foram dados no sentido de encurtar distância entre a Linguística – especialmente a aquisição da linguagem verbal oral – e a Ciência de Dados, com a abertura de novos horizontes de investigação.

REFERÊNCIAS

ANDRADE, A. J. P. **Ferramenta computacional para realização de testes de percepção e compreensão oral em crianças**. 2022. Projeto de Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Tecnológica Federal do Paraná, Santa Helena, 2022.

AREND, P. **Avaliação do desenvolvimento maturacional da linguagem em crianças de 4 a 6 anos em um município do oeste do Paraná**. 2021. Trabalho de Conclusão de Curso (Licenciatura em Ciências Biológicas) - Universidade Tecnológica Federal do Paraná, Santa Helena, 2021. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/27047. Acesso em: out. 2022.

BARTOSZECK, A. B.; BARTOSZECK, F. K. Neurociência dos seis primeiros anos: implicações educacionais. **Revista Educação**, p. 59-7, 2012. Disponível em: https://educacao.mppr.mp.br/arquivos/File/projeto_estrategico/argumentos_neurologicos_neurociencia_6_prim_anos_bartoszeck.pdf. Acesso em: out. 2022.

BRASIL. **Lei n. 13.709** – Lei Geral de Proteção de Dados. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: jun. 2023.

BUCHWEITZ, A.; MASON, R. A.; TOMITCH, L. M.; JUST, M. A. Brain activation for reading and listening comprehension: an fMRI study of modality effects and individual differences in language comprehension. **Psychol Neurosci**, v.2, p.111-23, 2009. DOI https://doi.org/10.3922/j.psns.2009.2.003. Acesso em: out. 2022.

BUHA, I.; FAMILI, A. Postprocessing in Machine Learning and Data Mining. **KDD2000PostWkshp**, 2000. Disponível em: https://www.kdd.org/exploration_files/KDD2000PostWkshp.pdf?fbclid=IwAR1D0420vP5L9b_jvKPc7JDBxuUN2XHraD_FChNfSF42EQggejm7_VcszPY. Acesso em: out. 2022.

CAGLIARI, L. C. **Análise fonológica**: introdução à teoria e à prática. Campinas: Mercado das Letras, 2002.

CÂMARA JR., J. M. **Problemas de lingüística descritiva**. 16.ed. Petrópolis: Vozes, 1997.

CÂMARA JR., J. M. **Estrutura da língua portuguesa**. 34.ed. Petrópolis: Vozes, 1996.

CASTANHEIRA, L. G. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. Belo Horizonte: UFMG, 2008.

COSTA, A. **Análise da produção técnico-científica dos docentes de duas jovens universidades de modelos distintos**: comparativo entre a Universidade Tecnológica Federal do Paraná (UTFPR) e a Universidade Federal do ABC (UFABC). 2019. Tese (Doutorado em Ensino de Ciência e Tecnologia) – Universidade Tecnológica Federal do Paraná, Ponta Grossa, 2019. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/4720/1/analiseproducaotecnicocientificadocentes.pdf#page=110&zoom=100,109,889. Acesso em: nov. 2022.

DAS, S. R. **Data Science**: Theories, models, algorithms, and analytics. Sanjiv Ranjan Das, 2016. Disponível em: https://srdas.github.io/Papers/DSA_Book.pdf. Acesso em: nov. 2022.

WEERASOORIYA, T. Role of the brain in perception and production of language. **Daily News**, 27 abr. 2021. Disponível em: https://www.dailynews.lk/2021/04/27/features/247558/role-brain-perception-and-production-language. Acesso em: jun. 2023.

DEHAENE, S. **Os neurônios da leitura**: como a ciência explica a nossa capacidade de ler. Trad. Leonor Scliar-Cabral. Porto Alegre: Penso, 2012.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v.39, n.11, nov. 1996. Disponível em: https://dl.acm.org/doi/pdf/10.1145/240455.240464. Acesso em: out. 2022.

FERNANDES, S.; VASILÉVSKI, V.; ARAÚJO, M. J. Um sistema computacional para suporte ao aprendizado da leitura via método fônico. In: Seminário de Extensão e Inovação, X. Seminário de Iniciação Científica e Tecnológica, XVV. 2020, Toledo/PR. **Anais...** Disponível em: https://eventos.utfpr.edu.br//sicite/sicite2020/paper/viewFile/7035/2139. Acesso em: nov. 2022.

FERREIRO, E.; TEBEROSKY, A. **Psicogênese da língua escrita**. Porto Alegre: Artmed, 1999.

FOLLADOR NETO, A.; SILVA, A. P.; YEHIA, H. C. Corpus CEFALA-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia / Corpus CEFALA-1: Audiovisual Database of Speakers for Biometric, Phonetic and Phonology Studies. **Revista de Estudos da Linguagem**, v. 27, n. 1, 2019. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/13378. Acesso em: out. 2022.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining**: Conceitos, técnicas, algoritmos, orientações e aplicações. 2.ed. Rio de Janeiro: Elsevier, 2015.

GOSWAMI, U. Principles of learning, implications for teaching: A cognitive neuroscience perspective. **Journal of Philosophy of Education**, v. 42, n. 3-4, p. 381-399, 2008.

GUIMARÃES, D. M. L. O. **Percurso de construção da fonologia pela criança**: uma abordagem dinâmica. 2008. Tese (Doutorado em Estudos Linguísticos) - Faculdade de Letras da Universidade Federal de Minas Gerais, Belo Horizonte. Disponível em: http://hdl.handle.net/1843/ARCO-7KVNSH. Acesso em: out. 2022.

HALL, M. A.; WITTEN, I. H.; FRANK, E. **The WEKA Data Mining Software**: An Update. SIGKDD Explorations, 2009. Disponível em: https://www.researchgate.net/publication/221900777_The_WEKA_data_mining_software_An_update. Acesso em: out. 2022.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and techniques. 3.ed. Walthan, MA (USA): Morgan Kauffman, 2011.

KINGHOST. **DBeaver**: Como acessar um banco SQL Server. s.data. Disponível em: https://king.host/wiki/artigo/acessar-sql-server-pelo-dbeaver/. Acesso em: out. 2022.

MACHADO, F. N. R. **Tecnologia e projeto de Data Warehouse**: uma visão multidimensional. 6.ed. São Paulo: Érica, 2013.

MANNING, C. D., SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. The MIT Press, 1999.

MICROSOFT. **Algoritmos de mineração de dados** (Analysis Services – Mineração de Dados). set. 2022. Disponível em: https://learn.microsoft.com/pt-br/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?source=recommendations&view=asallproducts-allversions. Acesso em: out. 2022.

MICROSOFT. **Algoritmos de aprendizado de máquina**. Disponível em: https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms. 2023a. Acesso em: jun. 2023.

MICROSOFT. **O que é o Power BI?** 2023b. Disponível em: https://powerbi.microsoft.com/pt-br/what-is-power-bi/ Acesso em: jun. 2023.

MOMENSOHN-SANTOS, T. M.; DIAS, A. M. N.; ASSAYAG, F. H. M. Processamento auditivo. In: MOMENSOHN-SANTOS, T. M.; RUSSO, I. **A prática da audiologia clínica**. 8.ed. São Paulo: Cortez, 2011. p. 275-290.

MULLER, A.; GUIDO, S. **Introduction to Machine Learning with Python**. O'Reilly Media, Incorporated, 2017.

MUSZKAT, M. Desenvolvimento e neuroplasticidade. In: MELLO, C.B. **Neuropsicologia do desenvolvimento**: conceito e abordagens. São Paulo: Memnon; 2005. p.26-45.

NEVES, O. **Implicações das neurociências na aprendizagem da leitura na pré-escola**. 2021. Trabalho de Conclusão de Curso (Licenciatura em Ciências Biológicas) – Universidade Tecnológica Federal do Paraná, Santa Helena, 2021. Disponível em: https://riut.utfpr.edu.br/jspui/handle/1/27041. Acesso em: ago. 2022.

NLTK PROJECT. **Documentation**. Natural Language Toolkit. 2022. Disponível em: https://www.nltk.org. Acesso em: out. 2022.

ORANGE. **Classroom Training**. 2022. Disponível em: https://orangedatamining.com/training/. Acesso em: jun. 2023.

PEREIRA, L. D. Introdução ao processamento auditivo central. In: ASSOCIAÇÃO BRASILEIRA DE FONOAUDIOLOGIA - ABA. **Tratado de audiologia**. 2.ed. São Paulo: GEN-Santos, 2011, p. 279-291.

RAMOS, A. **O desvio fonológico relacionado à consciência fonológica**. 2008. 58f. Monografia (Conclusão do Curso de Fonoaudiologia) – Feevale, Novo Hamburgo-RS, 2008.

ROB, P.; CORONEL, C. **Sistemas de banco de dados**: projeto, implementação e administração. Trad. da 8.ed. norte-americana. Cengage Learning, 2011.

R-PROJECT. The R Project for Statistical Computing. 2020. Disponível em: https://www.r-project.org. Acesso em: out. 2022.

SAID ALI, M. **Gramática secundária e Gramática histórica da língua portuguesa**. 3.ed. Brasília: Editora da UnB, 1964.

SCLIAR-CABRAL, L. **Sistema Scliar de Alfabetização**: Fundamentos. Florianópolis: Lili, 2013.

SCLIAR-CABRAL, L. **Guia prático de alfabetização**. São Paulo: Contexto, 2003.

SOUZA, N. R.; VASILÉVSKI, V. Acompanhamento das primeiras leituras na pré-escola. **Cadernos do SEI-SICITE**, 2022. No prelo.

SOUZA, I. **PostgreSQL**: saiba o que é, para que serve e como instalar. 2020. Disponível em: https://rockcontent.com/br/blog/postgresql/. Acesso em: out. 2022.

SHAYWITZ, S. E.; SHAYWITZ, B. A. Paying attention to reading: the neurobiology of reading and dyslexia. **Dev Psychopathol**, v.20, p.1329-49, 2008. Disponível em: https://www.cambridge.org/core/journals/development-and-psychopathology/article/abs/paying-attention-to-reading-the-neurobiology-of-reading-

and-dyslexia/1803914825F1FF08853F4D3EC723199A. Acesso em: out. 2022.

TAN, P.N.; STEINBACH, M.; KUMAR, V. Introdução ao Data Mining, Mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

THE INTERNATIONAL PHONETIC ALPHABET IN UNICODE – IPA. Disponível em: https://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm Acesso em: jun. 2015.

VASILÉVSKI, V.; ILIUK, J.; NEVES, O. M. A aprendizagem da leitura via método fônico na percepção dos pais. **Teoria e Prática da Educação**, v. 23, n.3, p. 56-76, set.-dez. 2020. Disponível em: https://periodicos.uem.br/ojs/index.php/TeorPratEduc/article/view/55742?fbclid=IwAR1nMBUEL1uwmC_NP6qV5UVbXVoSIHznVfrcxDBVuVyFVCr49JPzUBdIXDc?fbclid=IwAR1nMBUEL1uwmC_NP6qV5UVbXVoSIHznVfrcxDBVuVyFVCr49JPzUBdIXDc. Acesso em: out. 2022.

VASILÉVSKI, V.; ARAUJO, M. J.; BLASI, H. F. A Brazilian Portuguese Phonological-prosodic Algorithm Applied to Language Acquisition: A Case Study. WORKSHOP ON COGNITIVE ASPECTS OF COMPUTATIONAL LANGUAGE LEARNING (CogACLL), 5., 2014, Gothenburg, Sweden, **Proceedings**…, Chalmers University, Gothenburg, p.3-8. Disponível em: https://aclanthology.org/W14-0502.pdf. Acesso em: nov. 2022.

VASILÉVSKI, V. **Construção de um programa computacional para suporte à pesquisa em fonologia do português do Brasil**. 2008. 166f. Tese (Doutorado em Linguística) - Universidade Federal de Santa Catarina, Florianópolis. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/91849. Acesso em: abr. 2023.

WERTZNER, H. F. **Articulação**: aquisição do sistema fonológico dos três aos sete anos. 1992. Tese (Doutorado em Linguística) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/8/8139/tde-09022023-123509/publico/1992_HaydeeFiszbeinWertzner.pdf. Acesso em: jun. 2023.

WILLIAMS, J. **Teaching singing to children and young adults**. Compton Pub., 2012. Disponível em: https://www.jenevorawilliams.com/wp-content/uploads/2012/11/Inside-the-book.pdf. Acesso em: abr. 2023.

YONCHEVA, Y. N.; WISE, J.; MCCANDLISS, B. Hemispheric specialization for visual words is shaped by attention to sublexical units during initial learning. **Brain and Language**, v.145-146, p.23-33, June-July 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0093934X15000772. Acesso em: jun. 2023.