TRABALHO DE GRADUAÇÃO

Proposta de Sistema de Recomendação
com Aplicação de Técnicas de Clusterização
e Processamento de Linguagem Natural
de Descrições de Vinhos

Tainá Amorim Esteves

Brasília, Julho de 2019

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

Faculdade de Tecnologia

TRABALHO DE GRADUAÇÃO

Proposta de Sistema de Recomendação com Aplicação de Técnicas de Clusterização e Processamento de Linguagem Natural de Descrições de Vinhos

Tainá Amorim Esteves

Trabalho de Graduação submetido ao Departamento de Engenharia

Elétrica como requisito parcial para obtenção

do grau de Engenheira de Redes de Comunicação

Banca Examinadora

Prof. Georges Daniel Amvamve Nze, Dr., ENE/UnB _____
Orientador

Prof. Valério Aymoré Martins, MSc, ENE/UnB _____
Examinador Interno

Jorge Guilherme Silva dos Santos, MSc _____
Examinador Externo

FICHA CATALOGRÁFICA

ESTEVES, TAINÁ AMORIM

Proposta de Sistema de Recomendação com Aplicação de Técnicas de Clusterização e Processamento de Linguagem Natural de Descrições de Vinhos [Distrito Federal] 2019.

xvi, 45 p., 210 x 297 mm (ENE/FT/UnB, Engenheira, Engenharia Elétrica, 2019).

Trabalho de Graduação - Universidade de Brasília, Faculdade de Tecnologia.

Departamento de Engenharia Elétrica

1. Processamento de Linguagem Natural
2. K-means
3. Python
4. Georreferenciamento

I. ENE/FT/UnB

II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

ESTEVES T. A. (2019). *Proposta de Sistema de Recomendação com Aplicação de Técnicas de Clusterização e Processamento de Linguagem Natural de Descrições de Vinhos*. Trabalho de Graduação, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 45 p.

CESSÃO DE DIREITOS

AUTOR: Tainá Amorim Esteves

TÍTULO: Proposta de Sistema de Recomendação com Aplicação de Técnicas de Clusterização e Processamento de Linguagem Natural de Descrições de Vinhos.

GRAU: Engenheira de Redes de Comunicação ANO: 2019

Tainá Amorim Esteves

Depto. de Engenharia Elétrica (ENE) - FT

Universidade de Brasília (UnB)

Campus Darcy Ribeiro

CEP 70919-970 - Brasília - DF - Brasil

Agradecimentos

Primeiramente, agradeço a Deus por ter me dado força para concluir essa etapa da minha vida e por ter colocado pessoas tão especiais e queridas na minha vida. Dedico este trabalho aos meu pais maravilhosos e irmão, que sempre me incentivaram a estudar, me apoiaram e tiveram muita paciência comigo. Também agradeço ao meu namorado por ser sempre tão compreensivo, carinhoso e paciente todas as vezes em que precisei estudar ou quando saia muito triste ou muito feliz de alguma prova.

Também sou extremamente grata por todos os amigos que encontrei na graduação, estágios e projetos de pesquisa, pois fizeram meus dias mais leves e me ajudaram muito, inclusive meu amigo e colega de trabalho Jorge, que foi muito generoso ao me ajudar a solucionar dúvidas e dividir seus conhecimentos comigo.

Por fim e não menos importante, agradeço ao Professor Georges, que me orientou e apoiou minha ideia nada convencional de ter vinhos como um dos pilares do trabalho realizado, e as minhas amigas Gabriella e Kadichari, que foram essenciais para o rendimento que tive nesse semestre final.

Tainá Amorim Esteves

RESUMO

Este projeto tem como objetivo mostrar a implementação de um sistema de recomendação de vinhos por georreferenciamento, segundo sua variedade de uva. Isso se deu a partir da aplicação de técnicas de processamento de linguagem natural e um algoritmo de aprendizado de máquina voltado para clusterização. Durante a pesquisa, foram encontrados diversos estudos acerca de sistemas baseados nesse tipo de solução e chegou-se a conclusão de que o ramo da vitivinicultura seria melhor explorado, trazendo a atenção de potenciais consumidores.

ABSTRATCT

This project aims to show the implementation of a recommendation system by geo-referencing, according to its grape variety. This was based on the application of natural language processing techniques and a machine-learning algorithm for clustering. During the research, several studies were found about the systems based on this type of solution and the conclusion was reached that the branch of vitiviniculture would be better exploited, bringing the attention of potential consumers.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

AM	*Aprendizado de Máquina*
ML	*Machine Learning*
SIG	*Sistema de Informação Geográfica*
PLN	*Processamento de Linguagem Natural*
SGBD	*Sistema de Gerenciamento de Banco de Dados*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
JSON	*JavaScript Object Notation*
IA	*Inteligência Artificial*
NLTK	*Natural Language Toolkit*

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Atualmente, o Brasil ocupa a 14ª colocação no ranking de produção mundial de vinho e a 17ª colocação no ranking de consumo médio, segundo a Organização Internacional da Vinha e do Vinho [1]. Apesar disso, tanto a produção como consumo *per capita* no país ainda são pequenos, se comparado à outros países de produção de renome, como Itália, França e Chile, por exemplo, conforme mostrado nas Figuras 1.2 e 1.1.

Entretanto, o setor tem sido impulsionado nos últimos anos devido ao aumento da renda *per capita* no país, do número de consumidores de vinho [2], do volume de importações (Figura 1.3) e a subsequente redução dos preços de vinhos importados [3].

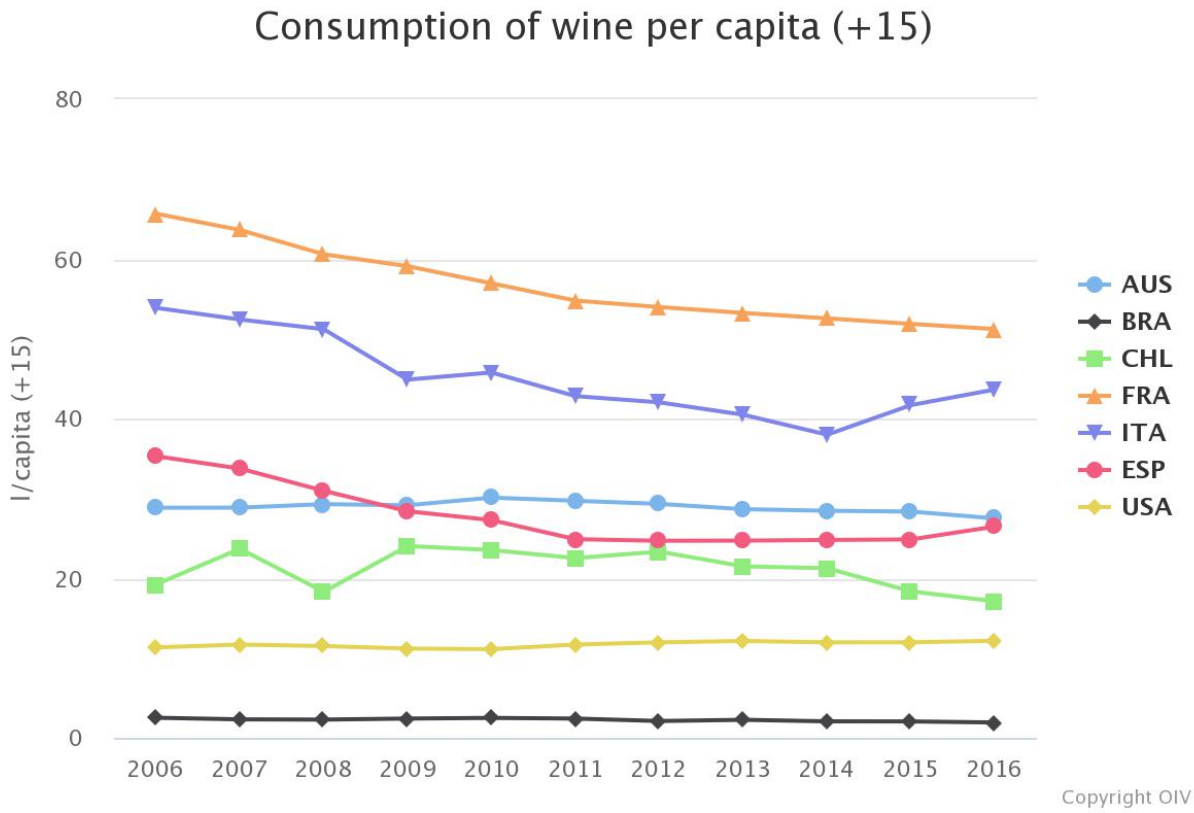

Figura 1.1: Consumo *per capita* de vinho de países em comparação com o Brasil entre 2006 e 2016 [1].

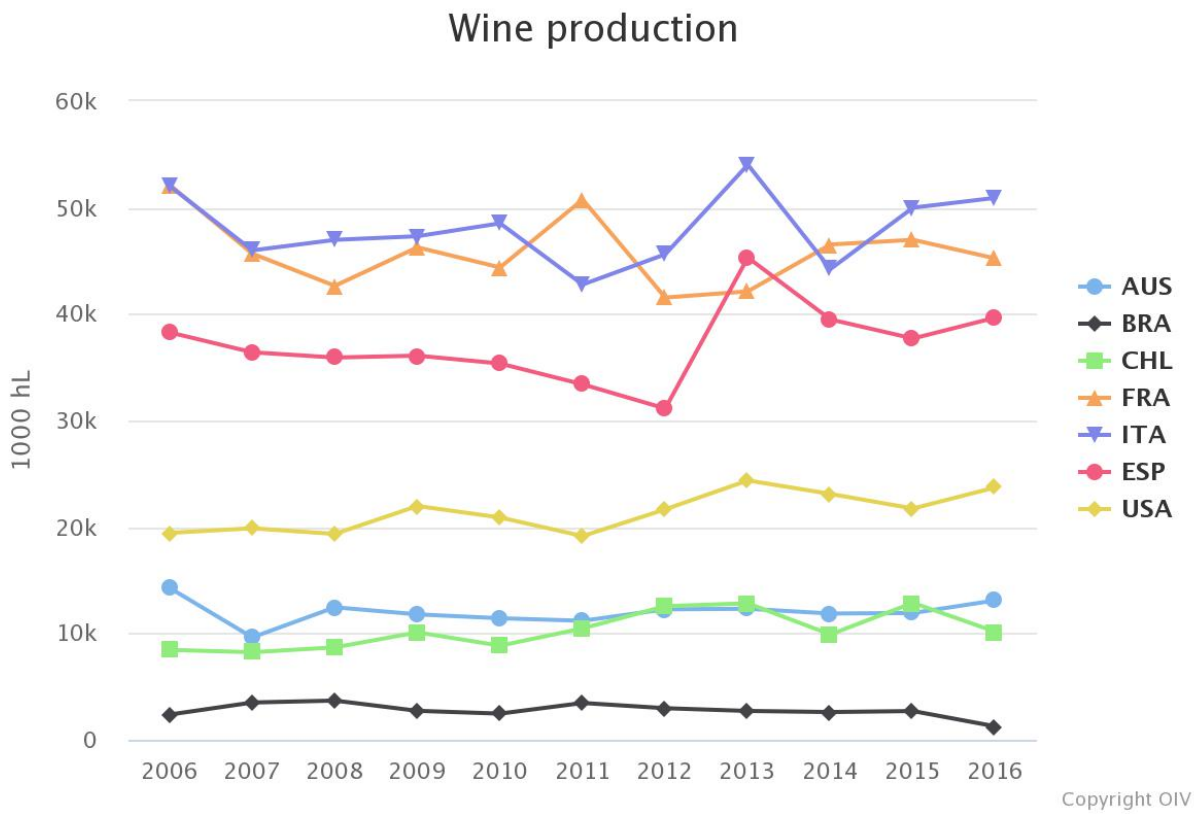

Figura 1.2: Produção de vinhos de países em comparação com o Brasil entre 2006 e 2016 [1].

Tais fatores tornam o setor extremamente receptivo à ferramentas que incentivem o consumo como, por exemplo, sistemas de recomendação [4]. Esse tipo de ferramenta, quando entrega sugestões adequadas ao perfil do consumidor, incentivam o consumo.

Porém, a análise de vinhos para sua categorização é extremamente custosa, visto que as características que descrevem os produtos são subjetivas, de modo que vinhos produzidos com a mesma uva podem apresentar características diferentes, dependendo do *terroir* e da subsequente qualidade da uva utilizada, por exemplo. Tais informações podem facilitar a análise das descrições de vinhos produzidos em conjunto com sua variedade e país de origem por meio processamento de linguagem natural e aprendizado de máquina.

É importante salientar também a importância das informações geográficas de origem das uvas, uma vez que uvas de uma mesma variedade e cultivas em regiões diferentes podem produzir vinhos com características diferentes. Um exemplo disso é o teor de açúcar, que é maior em regiões com maior incidência de luz solar.

Figura 1.3: Volume de importações de vinhos para o comércio brasileiro entre 2006 e 2016 [1].

1.2 OBJETIVO

Tendo em vista o potencial de crescimento de consumo gerado pela por uma ferramenta de recomendação de produtos eficaz e a dificuldade de se categorizar o produto da vitivinicultura, devido à subjetividade de suas características, o objetivo deste trabalho consistiu em aliar aprendizado de máquina, processamento de linguagem natural e técnicas de georreferenciamento para desenvolver uma aplicação que, a partir de uma seleção prévia de características feita pelo consumidor, gerasse informações úteis referentes a variedades de uva e país de origem, para auxiliá-lo na escolha de rótulos.

1.2.1 Objetivos específicos

Para realizar o desenvolvimento de tal aplicação, foram abordados os seguintes assuntos:

- Tratamento de *dataset* com relação a *outliers* e remoção de dados não utilizados.

- Aplicação de técnicas de processamento de linguagem natural para verificação das palavras mais frequentes das descrições de cada uma das variedades de uvas dos vinhos presentes no *dataset*.
- Aplicação do algoritmo *K-Means* para definição de *clusters* ou agrupamento de vinhos, a partir das palavras mais comuns nas descrições dos produtos presentes no *dataset*.
- Georreferenciamento das variedades mais semelhantes aos *clusters*, com mapas de densidade.
- Montagem de plataforma *Web* para coleta de dados do usuário e entrega de recomendação de regiões e variedades de uvas que serão mais apropriadas aos mesmos na hora de realizar a escolha de um produto.

Assuntos que não fazem parte da delimitação do tema foram citados como trabalhos futuros na seção 5.

1.3 ORGANIZAÇÃO DO TRABALHO

A partir dos próximos capítulos, o trabalho será dividido em quatro partes principais: fundamentação teórica, metodologia, resultados e conclusão. A fundamentação teórica está descrita no Capítulo 2 e aborda a fundamentação dos conceitos de Inteligência Artificial e georreferenciamento de dados usadas para o desenvolvimento do sistema proposto pelo trabalho.

A metodologia, apresentada no Capítulo 3, descreve as técnicas de Processamento de Linguagem Natural, os algoritmos de Aprendizado de Máquina e a construção da aplicação *Web*. O Capítulo 4, por sua vez apresenta o funcionamento do sistema e exemplos de resultados por ele entregues.

Por fim, foram apresentadas no Capítulo 5 as conclusões e possíveis aprimoramentos aplicáveis ao sistema desenvolvido.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 VITIVINICULTURA

A vitivinicultura pode ser definida como a união de práticas de vinicultura e viticultura [5], que são o conjunto de processos que têm como produto final o vinho [6] e o cultivo de vinhas [6], respectivamente. Tais práticas têm como objetivo a obtenção do vinho como produto final e as próximas seções trarão dados históricos e conceitos relevantes para o entendimento do assunto perante o trabalho.

2.1.1 Origem

As variedades de uvas das quais se têm conhecimento atualmente são distintas das uvas que eram encontradas na natureza antes da domesticação da videira *vitis vinífera*. O período de tal domesticação não tem data certa, tal como o início da produção de vinho, mas sabe-se que as primeiras videiras tiveram origem na área que vai desde a região central da Espanha até o mar Cáspio [7].

Porém, a partir de estudos arqueológicos em sementes de videiras e registros escritos, estima-se que tal domesticação da espécie e produção de vinho foi difundida pelos povos que habitavam às margens do Mar Mediterrâneo até popularização do consumo de vinho, no século VIII antes de Cristo [7], no Império Assírio, com os famosos banquetes celebrados nas classes mais altas da sociedade.

A partir disso, o hábito foi incorporado por outros povos da região, após contatos frequentes em transações comerciais, inclusive pela aristocracia grega, que transmitiu o hábito aos romanos, responsáveis por perpetuar o costume na região e levá-lo a outras regiões, conquistadas pelos mesmos. [8].

2.1.2 Situação atual

Com o consumo e produção de vinhos enraizados na população de regiões que têm origem no Império Romano e receberam influência da Igreja Católica, tais locais apresentam, até os dias atuais, grandes níveis de consumo e de produção dos mesmos. Contudo, com a globalização e crescente injeção de tecnologias na produção e cultivo de videiras, surgiram novos polos de produção de vinhos de qualidade em regiões fora do Velho Continente.

O fenômeno de globalização do vinho fomentou o crescimento do Enoturismo, que pode ser definido como *´´uma recente atividade caracterizada pelo deslocamento de pessoas a localidades que possuem tradição na produção de uvas e fabricação de vinhos, bem às regiões emergentes da atualidade. Durante o período dessas visitações, outros produtos e serviços são demandados e, consequentemente, oferecidos pela comunidade autóctone, gerando mais uma oportunidade à grandiosa indústria turística.*" [9].

2.1.3 Características do vinho

O vinho é um produto que tem grande variação de características, definidas a partir das variedades de uva que o compõem e do *terroir* de origem das mesmas [10], que é a combinação de quatro fatores que definem o ambiente no qual as vinhas crescem: clima, solo, terreno e cultura/-tradição.

Figura 2.1: Fatores que compõem o *terroir*. Adaptado de [11].

A combinação entre *terroir* e casta de uva geram produtos que podem ser classificados, quanto à classe, cor e teor de açúcar, de acordo com a legislação vitivinícola do Mercosul [12], e quanto à aspectos visuais, olfativos, gustativos e tácteis e observados em degustação dos vinhos [13]. Tal diversidade levam enólogos a descreverem os produtos de maneira subjetiva, de forma que a categorização a partir de dados referentes a características sensoriais seja uma oportunidade para aplicação de técnicas de inteligência artificial para realizar a interpretação dessas informações.

Além disso, a diversidade de *terroirs* dispostos nas diferente regiões vitiviníferas do mundo faz com que uma variedade de uva produza vinhos diferentes, de acordo com a sua região de origem, fazendo do georreferenciamento uma interessante ferramenta de análise para a variação de características de uma casta de uva de acordo com seu local de origem.

2.2 INTELIGÊNCIA ARTIFICIAL

Inteligência Artificial é o campo que engloba conhecimentos da ciência e das engenharias, que tem como objetivo a resolução de problemas de maneira rápida e eficiente com o auxílio de recursos computacionais. No entanto, tal definição é abrangente, visto que a IA tem diversas aplicações que vão desde entregar recomendações com base em comportamento, classificação de itens, detecção de fraudes, entre outros.

Para explicar o que este campo estuda, é necessário, primeiramente, definir o conceito de inteligência que, apesar de ser aplicado em outras áreas de estudo, pode ser definido como a capacidade de compreensão, aprendizado e armazenamento de informações. Quando o conceito de inteligência é associado ao poder de processamento de máquinas, é possível obter entidades artificialmente inteligentes [14].

2.2.1 Histórico

O nascimento Inteligência Artificial se deu a partir do trabalho de Warren McCulloh e Walter Pitts, de 1943, que propunha um modelo de neurônios artificiais, cujos estados - “ligado” e “desligado” - e organização determinariam seu comportamento [14]. Após este período, outros trabalhos e projetos relacionados foram apresentados, entretanto, o de maior destaque foi o teste de Turing, apresentado por Alan Turing em 1950.

2.2.1.1 Teste de Turing

O teste de Turing é aplicado até os dias atuais, por tratar de questões que são os alicerces da Inteligência Artificial. O teste consiste em um interrogatório, composto por três participantes - dois humanos e uma máquina - sendo que um dos humanos assume o papel de interrogador e mantém-se isolado dos outros dois participantes por uma barreira, conforme Figura 2.2.

O objetivo é verificar se a máquina a qual o teste é aplicado pode ser considerada um sistema de Inteligência artificial. Este é atingido se o interrogador, depois de realizar interações com os outros dois participantes, não conseguir distinguir se está se comunicando com a máquina ou com o humano entrevistados [15].

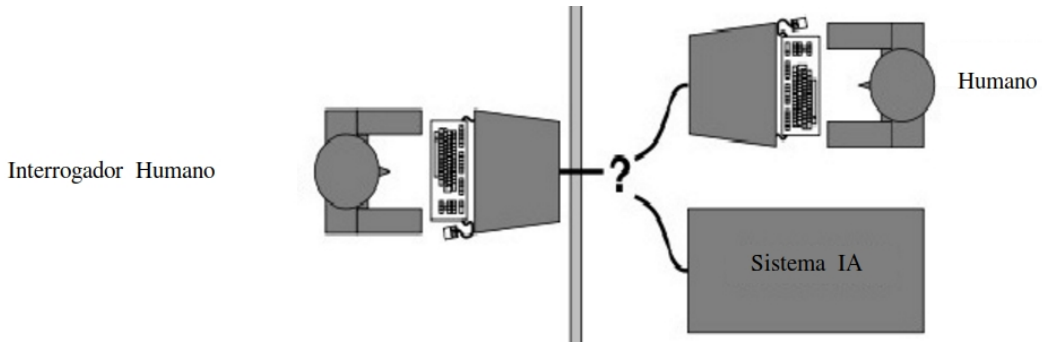

Figura 2.2: Ilustração da configuração dos participantes do Teste de Turing. Adaptado de [16].

Entretanto, para que a máquina possa confundir o entrevistador, ela precisa ter a capacidade e realizar as seguintes ações:

- processar a linguagem natural utilizada pelos humanos da interação;
- armazenar as informações obtidas e resgatá-las quando for necessário;
- tomar decisões ou realizar ações com base nas informações armazenadas;
- aprender para poder adaptar suas respostas caso não tenha o conhecimento armazenado.

As capacidades desejadas remetem a algumas das grandes áreas dentro do campo da Integência Artifical, como Processamento de Linguagem Natural e Aprendizado de Máquina.

2.2.2 Aprendizado de Máquina

A popularização de computadores pessoais, dispositivos móveis, como celulares e *tablets*, e a difusão de redes de comunicação sem fio possibilitaram um aumento significativo na criação de grandes volumes de dados, antes restritos a grandes centros de processamento [17]. O crescimento no volume de dados gerados todos os dias por usuários de dispositivos e aplicativos (mostrado na Figura 2.3), e a facilidade de coleta dos mesmos gerou, também, um aumento no uso de *Data Mining* aliado a técnicas de aprendizado de máquina por empresas, para auxiliar na tomada de decisões.

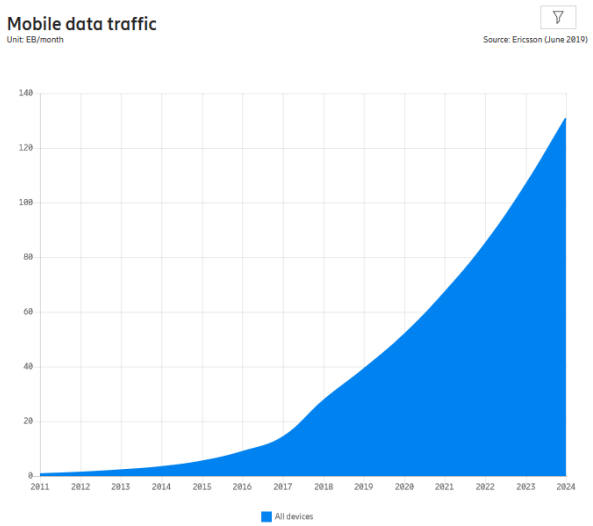

Figura 2.3: Crescimento do tráfego de dados móveis no período de 2011 a 2024 [18].

O aprendizado de máquina é o ramo de estudo da Inteligência Artificial que tenta reproduzir, a partir de algoritmos, os comportamentos de análise de informações, detecção de padrões e tomada de decisão utilizando-se de máquinas, com base em exemplos passados bem sucedidos, ou seja, aprender de maneira indutiva. O aprendizado indutivo pode ser: não supervisionado, por reforço e supervisionado.

Em algumas ocasiões, na literatura, é definido um quarto tipo de aprendizado indutivo, chamado de “semi-supervisionado”, um tipo de classificação especial em que são utilizados dados rotulados e não rotulados [19].

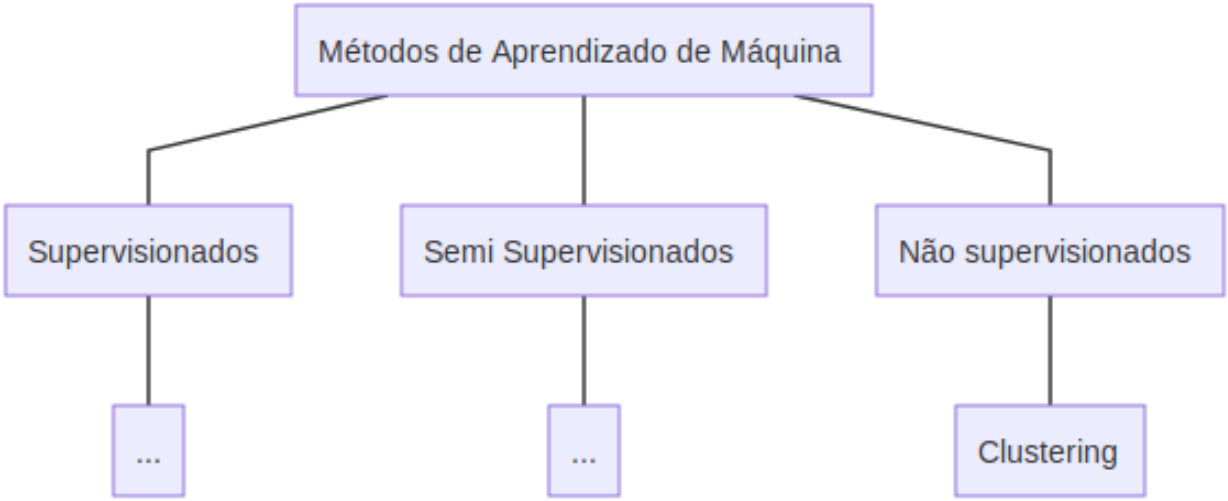

Figura 2.4: Métodos de aprendizado de máquina.

No primeiro caso, os dados são agrupados por similaridade e não são utilizados dados de treinamento para instruir como o modelo deve alocar cada entrada. Já no segundo, apesar de também não utilizar dados para treinamento do modelo, as saídas serão escolhidas de forma que resultem na obtenção da máxima recompensa [20]. No aprendizado supervisionado, por sua vez, é necessário que se tenha um *dataset* prévio para treinamento do modelo, que terá como saída, resultados previamente definidos.

2.2.2.1 Clusterização de dados

A clusterização de dados é uma das técnicas de análise de dados usadas em Aprendizado de Máquina que trata da organização das informações em grupos, utilizando-se de métricas, que variam de acordo com o algoritmo. Tal análise difere da classificação de dados porque, enquanto a primeira tem como objetivo definir os agrupamentos naturais de um conjunto de dados, na segunda, os grupos aos quais serão atribuídos os dados são pré-definidos, conforme visto na Figura 2.5 [21].

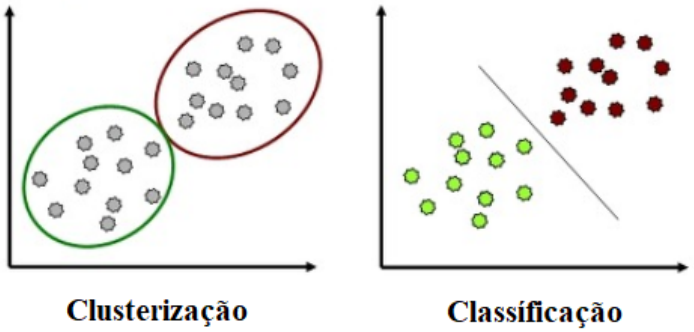

Figura 2.5: Representação das análises de dados com Clusterização e Classificação [22].

O *K-means* é um dos algoritmos de clusterização mais populares, principalmente por sua simplicidade [21]. Para aplicá-lo a um conjunto de dados, é necessário conhecer previamente o número de agrupamentos em que se deseja separá-lo. O agrupamento dos dados segue os seguintes passos [23]:

1. Com o número **K** de *clusters* informado, são estabelecidos os grupos iniciais de dados $C = \{c_k\}$ com $k \in \{1 \dots K\}$, a partir da média μ_j dos dados de cada *cluster*, são definidos seus respectivos pontos centrais iniciais.

2. Após a primeira definição de *clusters* e centróides, inicia-se uma iteração sobre cada uma das *i* amostras, na qual é feito o cálculo da soma das distâncias euclidianas ao quadrado J entre amostra e cada um dos centróides dado por [21]:

$$J(c_k) = \sum_{x_i \epsilon c_k} (|x_i - \mu_j|)^2. \tag{2.1}$$

3. Tais iterações são realizadas até que os *clusters* convirjam para a mesma organização, com J(C) assumindo o menor valor possível, e então é realizada a definição final dos *clusters* e seus centróides:

$$J(C) = \sum_{k=1}^{\mathbf{K}} \sum_{x_i \epsilon c_k} |x_i - \mu_j|^2. \tag{2.2}$$

2.2.3 Processamento de Linguagem Natural

A linguagem natural é o meio com o qual os seres humanos se comunicam. A comunicação, apesar de ser inata, varia de acordo com o idioma e a região dos participantes, visto que existem diferenças entre grupos sociais e diferentes regras para cada idioma. Ao trazer tal conceito para a área de Inteligência Artificial, obtém-se a definição de Processamento de Linguagem de Natural - processo de compreensão, análise e geração de informação na forma textual feito por máquinas [24].

O crescimento do volume de dados nos últimos anos [18] acarretou, também, em um aumento na quantidade informações textuais disponíveis para análise em diversas áreas e, consequentemente, o desenvolvimento de técnicas de PLN para realizar a mineração de textos, utilizada para extração de informações contidas em base de dados textuais [25].

O PLN é dividido em etapas relacionadas às áreas da linguística de sintaxe, semântica e pragmática [26]. A primeira está relacionada à concordância das palavras num texto, a segunda, ao significado individual de cada uma e a pragmática, por sua vez, diz respeito ao contexto em que o texto está inserido [6].

Como é possível verificar na Figura 2.6, PLN envolve diversos estágios. A primeira etapa a ser realizada em PLN é o pré-processamento do texto a ser analisado, executado a partir da conversão de um texto em formato de dados brutos, representado por uma sequência de *bits* em unidades linguísticas significativas [27].

O pré-processamento pode ser dividido em dois estágios: triagem de documentos e segmentação de texto. O primeiro estágio se refere à conversão de arquivos em textos bem definidos, e tal processo contém passos como identificação de codificação de caractere do texto, do idioma utilizado na escrita e remoção de elementos não textuais. A segmentação de texto, por sua vez, é a separação do texto em unidades menores, como sentenças e palavras. A segmentação em palavras é feita a partir da quebra de uma sequência de caracteres, feita com o uso de caracteres de pontuação e espaço em branco como delimitação, gerando *tokens* [27]. A segmentação de texto pode ser mencionada também com o termo *tokenização*, adaptação não oficial de *tokenization* e forma escrita utilizada em alguns trabalhos acadêmicos.

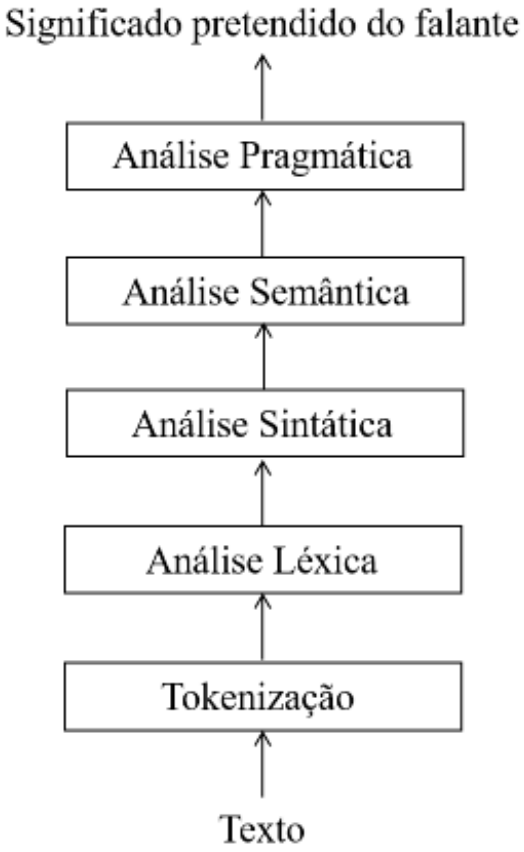

Figura 2.6: Estágios de análise em PLN. Traduzido de [26]

.

Após a *tokenização* do texto, é feita a remoção das chamadas *stop words*, isto é, palavras que não acrescentam informação relevante ao texto. Em seguida, deve ser realizada a análise léxica dos *tokens*, passando por processo de lematização e radicalização, definidos como a transformação dos verbos em suas formas no infinitivo e a redução das palavras às suas raízes morfológicas, respectivamente [28].

Os estágios seguintes da análise PLN tratam de unidade de informação maiores do que palavras, como frases e textos. Como o trabalho se limitou a *tokenização* e análise léxica, tais temas não serão abordados nessa seção.

2.3 GEORREFERENCIAMENTO DE DADOS

2.3.1 Sistemas de Informação Geográfica

O georreferenciamento de dados pode ser descrito como a associação de informações a um ponto, linha ou área de um mapa, e se faz importante quando tais informações são decisivas para análise de dados a ser realizada. Os sistemas automatizados para analisar e fazer operações com esses dados, de forma a gerar conhecimento ou dar suporte a tomada de decisão são denominados Sistemas de Informação Geográfica (SIGs), que realizam captura, preparação, administração, análise, manipulação e apresentação dos dados [29].

2.3.2 SIGs e a Vitivinicultura

Dados geográficos da vitivinicultura são utilizados para análise de variação nas áreas de estudo de definição de local viável mais adequado para cultivo de vinhas e desenvolvimento de técnicas para agricultura de precisão e identificação de vinhedos [30]. Um SIG auxilia na definição do local ao possibilitar a análise das múltiplas características dos *terroirs* que influenciam na qualidade do vinho, na decisão pelo local no qual essas características serão otimizadas e no gerenciamento das informações após o início do cultivo e produção do vinho [31].

2.4 WEB E APLICAÇÃO

Nessa seção serão tratados conceitos importantes para o desenvolvimento de soluções informatizadas de processamento, armazenamento e apresentação de dados.

2.4.1 Modelo de múltiplas camadas

O modelo de múltiplas camadas é uma técnica de setorização de tarefas que ganhou força com o aumento da complexidade de aplicações [32]. Um sistema construído a partir de uma arquitetura de múltiplas camadas possui, em sua forma mais básica, três camadas principais: apresentação, domínio e base de dados.

A camada de apresentação tem como atribuição tratar das interações entre usuário e aplicação, que podem ser provisionamento de serviços, entrega de informações e respostas a solicitações [32]. A camada de domínio, por sua vez, trata da lógica do sistema, que inclui autenticação de usuários e cálculos acerca de dados inseridos pelos usuários e dados armazenados. Já a camada de base de dados é a responsável pela comunicação de outros sistemas detentores de informações para o funcionamento da aplicação.

2.4.2 Python

Python é uma linguagem de programação interpretada, de alto nível e orientada a objeto, projetada por Guido Van Rossum em 1990 [33].É muito utilizada atualmente em computação científica devido a uma série de características desejáveis para processamento e análise de dados [34], entre elas:

- possui licença *Open Source*, o que possibilita que sejam criadas e distribuídas aplicações feitas com *Python* livremente [34];
- é compatível com as mais diversas plataformas de sistemas operacionais [34];
- a sintaxe é elegante e simplificada [35];
- há uma enorme quantidade de bibliotecas, que contêm tipos de dados, funções, exceções e módulos [35], aumentando a capacidade do programador de realizar diversas tarefas [36].

Entre as bibliotecas *Python* existentes, estão algumas que devem ser citadas devido ao seu uso em desenvolvimento de ferramentas de ciência de dados:

1. Pandas: provê estruturas de dados simplificadas para facilitar a manipulação de *datasets* e ferramentas de análise de dados, em memória [37].
2. *NLTK*: provê funções específicas para aplicação de técnicas de NLP [38].
3. *Numpy*: provê objetos em forma de *array* N-dimensional e funções que realizam operações com tais objetos [39].
4. *SQLAlquemy Toolkit and Object Relational Mapper*: conjunto de ferramentas utilizado para realizar conexão e operações com bancos de dados usando *Python* [40].

5. *Scikit-Learn*: conjunto de ferramentas que implementam algoritmos de ML supervisionados e não-supervisionados [41].

2.4.3 Bancos de Dados

Tecnologias de armazenamento de dados como bancos de dados foram um dos fatores que influenciaram no crescimento do uso de computadores, visto que o armazenamento e processamento de informações desempenham um papel fundamental no desenvolvimento das mais diversas aplicações computacionais [42].

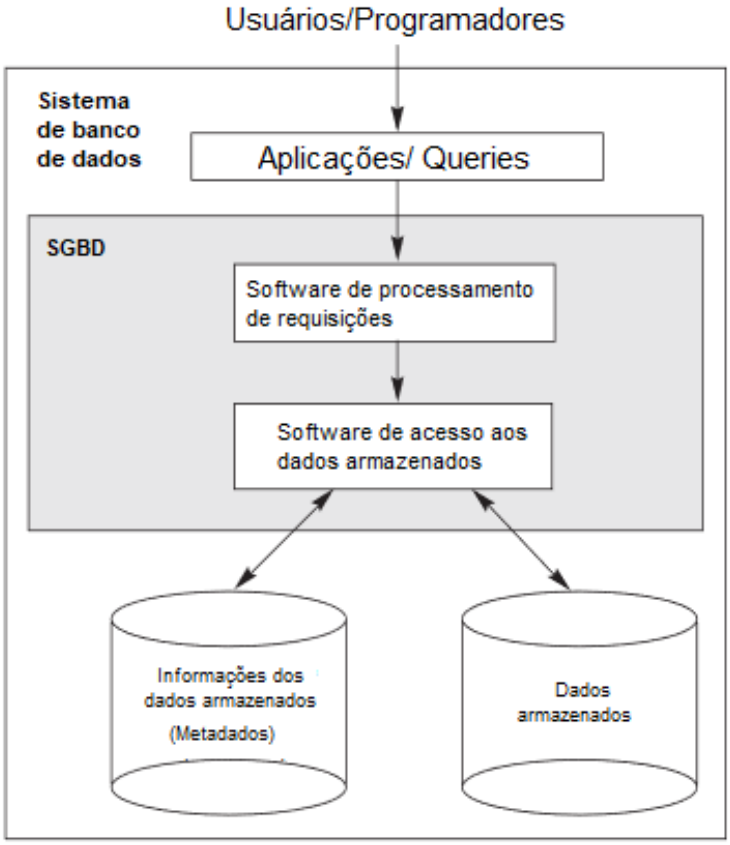

Figura 2.7: Ambiente simplificado de um Sistema de Banco de Dados. Traduzido de [42]

.

É possível definir um banco de dados como um conjunto de dados no qual toda a informação é organizada e relacionada [42, 43] e os Sistemas de Gerenciamento de Banco de Dados (SGBDs) como *softwares* que controlam o armazenamento, organização e consulta aos bancos de dados, conforme ilustrado na figura 2.7. Um dos SGBDs de código aberto utilizados atualmente é o *MySQL*. Ele é classificado como relacional, tendo em vista que os dados são organizados em tabelas, de acordo com a relação entre eles [44]. A junção de banco de dados e SGBD configura um sistema de banco de dados [42].

3 METODOLOGIA E IMPLEMENTAÇÃO

Este capítulo irá apresentar a estrutura do sistema desenvolvimento, bem como as decisões tomadas em cada etapa e explanações sobre as implementações realizadas.

3.1 ARQUITETURA DO SISTEMA

O sistema proposto foi desenvolvido seguindo o modelo de aplicações *web* de três camadas, no qual há uma camada de apresentação, uma camada de domínio e a base de dados conforme mostrado na Seção 2.4. A ferramenta escolhida para atuar na camada de base de dados foi o *SGBD MySQL*, instalado em ambiente *Ubuntu 18.04*. A camada de domínio, aplicada ao sistema proposto, engloba pré-processamento de dados, aplicação de algoritmo de NLP nos dados pré-processados, exportação de modelo de clusterização, envio de informações dadas pelo usuário e resposta do sistema com a correspondente informação de recomendação de países e variedades de uvas. E a camada de aplicação, por sua vez, refere-se a componentes do sistema com os quais o usuário interage, no caso, o sistema *web*.

Figura 3.1: Fluxograma do sistema.

3.2 PROCESSAMENTO

A parte inicial do trabalho englobou as tarefas de escolha de um conjunto de dados que contivesse os dados adequados para aplicação de PLN, tratamento dos dados escolhidos e aplicação das técnicas de PLN descritas na Seção 2.2.3. Cono pode ser visto na Figura 3.2, o processador PLN busca os dados pré-processados e armazenados no banco de dados e coordena o processador *K-means* e o relacionador para o processamento de dados. As tarefas atribuídas a cada entidade do sistema serão explicadas nas seções seguintes.

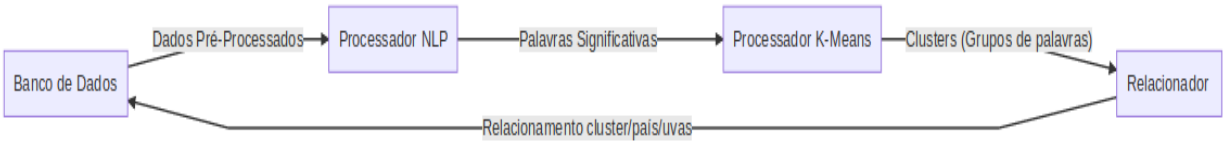

Figura 3.2: Diagrama de sequência do processamento de dados do sistema proposto.

3.2.1 Definição de base de dados

O *dataset* utilizado como fonte de para a aplicação dos algoritmos e técnicas deste trabalho foi retirado da plataforma *Kaggle* [45] e foi escolhido por conter dados de país de origem, descrição e variedade de uvas usadas na fabricação de cerca de cento e trinta mil vinhos. O autor do *dataset* coletou os dados em inglês a partir da página *Web Wine Enthusiast* por meio de técnicas de *Web scraping* [46] e o *dataset* resultante, utilizado como ponto de partida deste projeto, foi organizado conforme estrutura apresentada na figura 3.3.

country	description	designation	points	price	province	region_1	region_2	taster_name	ster_twitter_hand	title	variety	winery
Italy	Aromas include trop…	Vulkà Bianco	87	nan	Sicily & Sardinia	Etna	nan	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco…	White Blend	Nicosia
Portugal	This is ripe and fruity, …	Avidagos	87	15	Douro	nan	nan	Roger Voss	@vossroger	Quinta dos Avidagos 201…	Portuguese Red	Quinta dos Avidagos
US	Tart and snappy, the …	nan	87	14	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot G…	Pinot Gris	Rainstorm
US	Pineapple rind, lemon …	Reserve Late Harvest	87	13	Michigan	Lake Michigan Shore	nan	Alexander Peartree	nan	St. Julian 2013 Reserve…	Riesling	St. Julian
US	Much like the regular bott…	Vintner's Reserve Wild…	87	65	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner…	Pinot Noir	Sweet Cheeks
Spain	Blackberry	Ars In Vitro	87	15	Northern	Navarra	nan	Michael	@wineschach	Tandem 2011	Tempranillo-	Tandem

Figura 3.3: Organização dos dados do *dataset* utilizado.

3.2.2 Pré-processamento de dados do dataset

O *dataset* escolhido, apesar de conter as informações desejadas para o trabalho, contém dados dispensáveis à proposta, como preço, regiões de origem, pontuação e nome do provador do vinho, assim como o identificador do usuário no *Twitter*. Além disso, foi verificado que existem 11131 vinhos com títulos duplicados, 10016 descrições duplicadas, um vinho sem informação referente a casta de uva utilizada para sua fabricação e 63 vinhos sem a informação do país de origem no *dataset*.

Tendo isso em vista, foi feito o pré-processamento do *dataset* com um interpretador *Python* e a biblioteca Pandas, no qual os dados foram armazenados em um *dataframe*, para

- remoção das colunas *designation*, *region_1*, *region_2*, *taster_name*, *taster_twitter_handle*, *price*, *points* e *province* com o método *DataFrame.drop*;

- identificação e remoção de entradas com valores na coluna *title* duplicados, com os métodos *Dataframe.duplicate* e *Dataframe.drop_duplicate*;
- remoção das linhas em que faltavam informações referente às colunas *country*, *title* e *variety*, com os métodos *Dataframe.dropna*.

Após a realização do procedimentos acima, os dados contidos no *dataframe* foram escritos em uma tabela no banco de dados da aplicação, cuja estrutura é mostrada na tabela 3.1.

id	country	description	title	variety	winery
...	...	...	...	...	...

Tabela 3.1: Descrição da tabela *MySQL* com os dados pré-processados.

3.2.3 Processamento de Linguagem Natural

A aplicação de técnicas de PLN no trabalho desenvolvido tem como objetivo identificar quais palavras usadas na descrição dos vinhos são realmente importantes para a análise das características visuais, gustativas, olfativas e tácteis que cada variedade de uva atribui a um vinho, baseando-se em um modelo de análise que atribui valores de importância a cada palavra na descrição em que se encontra [47]. Para isso, os dados pré-processados foram coletados no banco de dados e armazenados em um *dataframe*, para que os dados referentes à coluna *description* fossem armazenados em um *array Numpy*, formando uma coleção de textos. Em seguida, foram estabelecidos parâmetros para *tokenização*, radicalização e realizada a análise de frequência das palavras, detalhados nessa seção.

Para a *tokenização* das descrições dos vinhos e radicalização de palavras, descritos na Seção 2.2.3, foram utilizados os módulos *RegexpTokenizer* e *SnowballStemmer*, sendo ambos da biblioteca NLTK. O primeiro dividiu cada *string* em *substrings* de acordo com as expressões regulares que foram escolhidas, de forma que, aplicado ao *array* de descrições, cada descrição se tornou um *array* de palavras que a compunham [38]. O segundo reduziu todas as palavras aos seus respectivos radicais na lingua inglesa que, como definido na Seção 2.2.3, é a parte da palavra que exprime seu significado básico [38].

Após geração dos *tokens*, definiu-se duas listas: *desc*, para armazenar as descrições dos vinhos, e *words*, que foi utilizada para armazenar as expressões mais recorrentes em *desc* e que contém os mil tokens mais frequentes, considerando a remoção das *stop words*. Assim, o módulo *TfidfVectorizer* [48] pôde ser utilizado para montar uma matriz *X* que referenciasse as duas listas supracitadas.

O módulo *TfidfVectorizer* é utilizado em textos para calcular a relevância de cada palavra presente no mesmo, de forma que quanto mais recorrente é um *token* nele (após a remoção de *stop words* e radicalização), maior é seu peso no texto. Assim, a matriz *X* passa a ser formada por três colunas: posição na lista *desc*, posição na lista *words* e fator de significância [48]. Assim, essa matriz é representa o fator de significância de cada palavra e sua respectiva posição na descrição.

A estrutura da matriz *X* é apresentada na Figura 3.4b, onde é possível observar, em destaque, que *token* número 5 do vetor *words* (3.4a) está presente tanto na descrição 0 (Figura 3.5) quanto na descrição 70198 (Figura 3.6). Entretanto, nas duas descrições, o mesmo *token* possui diferentes fatores de significância que, posteriormente, foram utilizados como métrica para clusterização.

```
This has great depth of flavor with its fresh apple and pear fruits and touch of
    spice. It's off dry while balanced with acidity and a crisp texture. Drink now.
```

Figura 3.5: Exemplo de Descrição.

```
A dry style of Pinot Gris, this is crisp with some acidity. It also has weight and a
    solid, powerful core of spice and baked apple flavors. With its structure still
    developing, the wine needs to age. Drink from 2015.
```

Figura 3.6: Exemplo de Descrição.

0	str	1	absolut
1	str	1	abund
2	str	1	accent
3	str	1	access
4	str	1	accompani
5	str	1	acid
6	str	1	acr
7	str	1	ad
8	str	1	add
9	str	1	addit
10	str	1	afford

(a) Exemplo de Vetor *words*.

```
In [21]: print (X)
  (0, 893)      0.22224870229893656
  (0, 811)      0.38624471826425216
  (0, 486)      0.24791811415032405
  (0, 332)      0.3699716572426941
  (0, 727)      0.33718657381997785
  (0, 252)      0.2709973444644745
  (0, 382)      0.2246542174803401
  (0, 626)      0.2573731998393131
  (0, 205)      0.20213041443533675
  (0, 5)        0.13489635811643874
  (0, 837)      0.3489644254861687
  (0, 846)      0.34409792737881123
  (1, 727)      0.28997904576881645
  (1, 626)      0.22133987735738556
  (1, 472)      0.19272231125941386
  (1, 630)      0.31714217769613995
  (1, 587)      0.20815440423366327
  (1, 81)       0.27132625015495465
  (1, 842)      0.2606085909271817
  (1, 598)      0.11550868804065288
  (1, 71)       0.20162706120698798
  (1, 586)      0.2748436194146044
  (1, 418)      0.23327564876933363
  (1, 392)      0.3369133916885273
  (1, 513)      0.286059557574059
  :     :
  (70197, 364)  0.2927472254704489
  (70197, 130)  0.234753971133331
  (70197, 168)  0.27871140518048454
  (70197, 705)  0.2843404104262782
  (70197, 220)  0.2823679874222218
  (70197, 413)  0.26612594750338675
  (70197, 520)  0.3348878137443535
  (70197, 181)  0.33080498615473464
  (70198, 205)  0.2106196226053761
  (70198, 5)    0.14056182547635937
  (70198, 28)   0.20932123907345276
  (70198, 830)  0.16768225478869972
  (70198, 257)  0.1587763237875482
  (70198, 860)  0.20290075445908692
```

(b) Exemplo da organização da matriz resultante do TF-IDF.

Figura 3.4: Exmplos do vetor *words* e da matriz TF-IDF.

3.3 CLUSTERIZAÇÃO

Para a aplicação do *K-means* na matriz obtida com o *TfidfVectorizer*, foi definido que o número de *clusters n_clusters* deveria ser igual ao de variedades de uvas dos vinhos, visto que desejava-se obter a relação da presença das palavras mais frequentes às descrições dos vinhos de cada variedade. Além disso, foi estabelecido o que o número *n_init* de vezes que o algoritmo iria com diferentes centroides iniciais escolhidos aleatoriamente, para se obter as palavras correspondentes aos centroides de cada *cluster*, seria igual a dez (quantidade de variedades de uvas diferentes presentes no *dataser*).

Após rodar o algoritmo e obter os *clusters* desejados, foram obtidas, a partir do atributo *clusters_centers_*, as dez palavras mais próximas aos centroides de cada *cluster* de vinhos, visto que a menor distância ao centroide é a medida de similaridade utilizada pelo *K-means*. Tais informações foram, então, armazenadas em um dicionário do *Python* para análise posterior, descrita na Seção 3.4.1.

Além disso, foi feita a atribuição de uma pontuação conforme a contagem de ocorrência da tupla *(cluster, país, variedade)* na base de dados. Essa pontuação foi utilizada, posteriormente, como métrica de recomendação de país de origem e variedade de uva, para escolha de vinho de acordo essas características.

Tais tarefas configuraram a criação do modelo de clusterização que foi utilizado nas etapas seguintes do trabalho. No entanto, a aplicação, tanto das técnicas de PLN como do algoritmo *K-means*, foram extremamente demoradas (com duração superior à trinta minutos) e exigiu uma capacidade de processamento considerável. Surgiu, então, a necessidade de persistência do modelo criado, de modo que esse não precisasse ser rodado toda vez que fosse necessária a obtenção de dados relacionados aos *clusters* obtidos.

O módulo *pickle* do *Python* foi utilizado para a exportação do modelo em um arquivo que foi importado posteriormente e utilizado na função de predição de *cluster*. Tal função recebe uma *string* e aplica o método *Predict*, que prevê o *cluster* que mais se aproxima dos dados contidos nela.

3.4 APLICAÇÃO WEB

Nesta etapa do trabalho, foram desenvolvidas a interface que recebe os dados referentes às preferências do usuário e retorna a recomendação calculada pelo sistema, a lógica de programação responsável por gerar a recomendação para o usuário e a renderização das informações contidas na recomendação de maneira georreferenciada. As interações entre usuário e funcionalidades do *software* estão representadas na Figura 3.7.

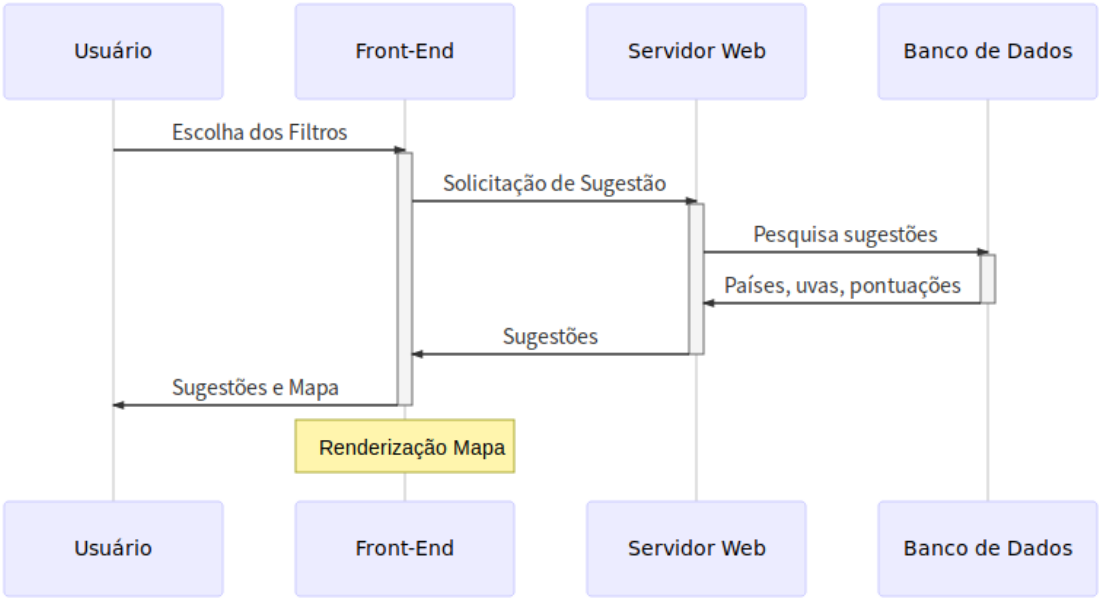

Figura 3.7: Diagrama de sequência da aplicação Web do sistema proposto.

3.4.1 Back-end

No sistema proposto, o *back-end* é responsável por trabalhar as informações de preferência do usuário, aplicá-las à função de predição de *cluster* e, com base no *cluster* calculado, retornar as quatro opções correspondentes de país e variedade de uva com maior pontuação de acordo com a métrica definida na Seção 3.3.

Para montar o filtro de opções referentes às características olfativas, gustativas, tácteis e visuais usadas para classificar vinhos, foi necessário realizar uma análise manual dos dados das expressões (*tokens*) mais comuns relacionadas a cada *cluster*.

O primeiro passo dessa análise foi pesquisar o significado pragmático das expressões e organizá-las de acordo com as características a que elas se referem, na plataforma *WineFrog* [49]. Por exemplo, expressões como "*sugar*"(açúcar) e "*dry*"(radical da palavra seco em inglês) remetem a vinhos com alto e baixo teor de açúcar, respectivamente. O conjunto de expressões relacionados a cada característica foi armazenado em seu respectivo dicionário e então, cada dicionário foi relacionado com sua opção correspondente do filtro montado no *front-end* da aplicação.

Após a realização da análise inicial dos dados dos *clusters*, o sistema foi preparado para receber a requisição do usuário em formato *JSON* (Figura 3.8a) e construir uma *string* a partir da correspondência entre os parâmetros contidos na requisição e os dicionários relacionados, visto que a função de predição desenvolvida na etapa descrita na seção 3.3 recebe como entrada uma *string*.

Em seguida, foi realizado o cálculo do *cluster* e a consulta na tabela do banco de dados referentes às informações de pontuação, variedade de uva e país de origem, relacionadas ao *cluster* calculado. Feito isso, o sistema devolve uma resposta *HTTP* ao *front-end*, em formato *JSON*, contendo as quatro variedades de uva com maior pontuação e seus respectivos países, conforme mostrado na Figura 3.8b.

```
{
    "red": true,
    "white": false,
    "rose": false,
    "sweet": false,
    "dry": false,
    "light": false,
    "medium": false,
    "high": false,
    "tannin": false,
    "acid": false,
    "fruit": false,
    "mineral": false,
    "dark": false,
    "yellow": false,
    "young": false,
    "old": false,
    "complex": true,
    "intermediary": false,
    "simple": false
}
```

(a) Exemplo de mensagem da requisição HTTP que solicita recomendação.

```
[{
    "country": "US",
    "points": 136.438798661134,
    "variety": "Bordeaux-style Red
        Blend"
}, {
    "country": "Italy",
    "points": 92.7841460800109,
    "variety": "Red Blend"
}, {
    "country": "France",
    "points": 18.5856441682009,
    "variety": "Bordeaux-style Red
        Blend"
}, {
    "country": "Spain",
    "points": 6.91558852770267,
    "variety": "Red Blend"
}]
```

(b) Exemplo de mensagem da resposta HTTP que solicita recomendação.

Figura 3.8: Requisição e Resposta

3.4.2 Front-end

O *front-end* da aplicação desenvolvida é responsável por receber os dados de preferência do usuário, apresentar a resposta do sistema. A interface *web* consiste em uma caixa de *checkboxes* com as opções referentes às características que o usuário gosta/deseja em um vinho e o mapa onde serão mostrados os países de origem das uvas que são mais compatíveis com as escolhas do usuário, conforme mostrado na Figura 3.9.

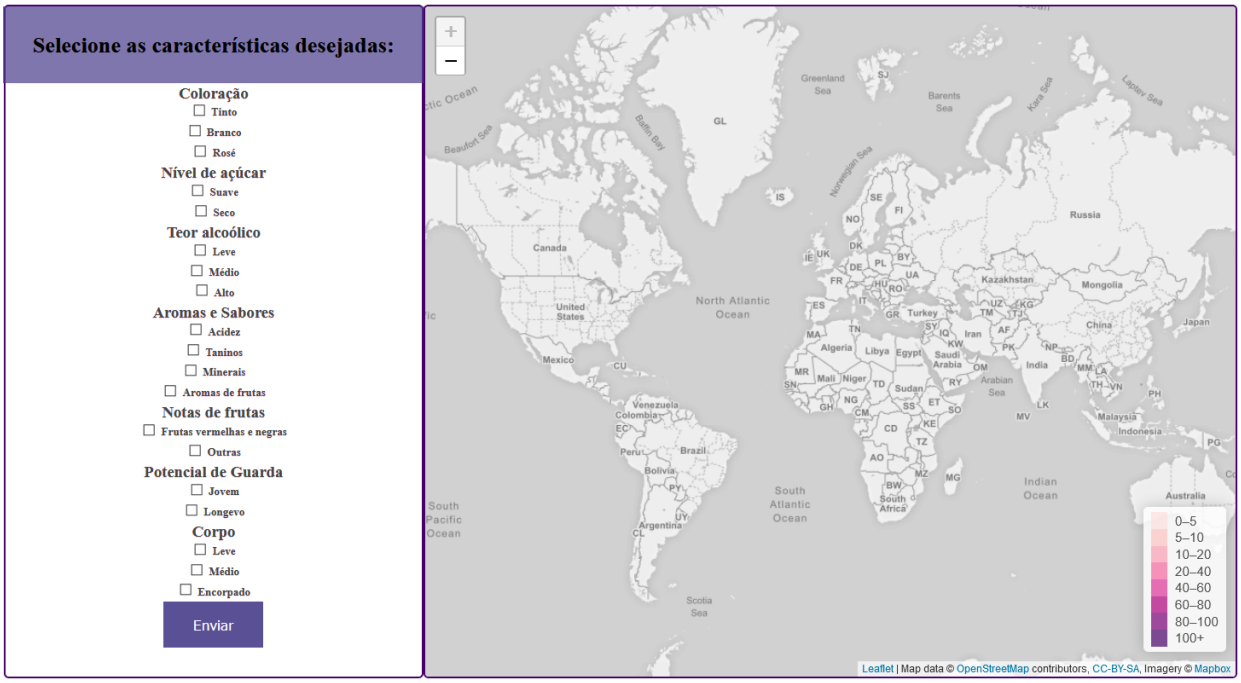

Figura 3.9: Visão inicial da interface *Web* da aplicação antes da seleção e requisição do usuário.

Conforme apresentado na Seção 3.4.1, a interface envia os dados submetidos pelo usuário em formato *JSON* por meio de uma requisição *HTTP* e recebe, no mesmo formato, as informações que devem ser mostradas no mapa. A apresentação das informações é feita com sua renderização no mapa, de forma que o mesmo se torne um mapa coroplético, usando a intensidade da cor que preenche a área do país referente à variedade de uva para representar seu nível de indicação. A adição desses mapas foi feita com o auxilio da *Leaflet*, uma biblioteca *JavaScript open source* de mapas interativos [50] e, na Seção 4, serão apresentados exemplos de funcionamento da aplicação e diferentes exibições do mapa.

4 RESULTADOS

4.1 INFORMAÇÕES DOS CLUSTERS

O processamento das informações relacionadas às descrições dos vinhos resultou em uma estrutura contendo o identificador de cada *cluster* e os dez *tokens* mais frequentes às descrições dos vinhos de cada agrupamento, conforme apresentado na Tabela 4.1. Tais informações foram utilizadas para validar o modelo construído, de forma que um *cluster* que contém os *tokens* "*black*", "*tannin*" e "*blend*" terá maior correlação com vinhos tintos de uvas como *Cabernet Sauvignon* e *Petit Verdot*, que produzem vinhos com mais taninos, que têm aromas de frutas vermelhas como mirtilo e amoras, por exemplo.

Cluster	Tokens relacionados									
0	blackberri	currant	tannin	oak	cabernet	rich	chocol	ripe	dri	tannic
1	pinot	noir	cherri	raspberri	silki	cola	acid	dri	rich	oak
2	fruiti	acid	crisp	light	fresh	fruit	red	textur	soft	ripe
3	wood	age	fruit	rich	ripe	tannin	structur	acid	black	spice
4	berri	plum	herbal	palat	fruit	oak	earthi	red	spice	oaki
5	syrah	grenach	blend	fruit	sirah	petit	cherri	black	red	pepper
6	cranberri	tart	cherri	palat	fruit	light	red	raspberri	spice	dri
7	medium	bodi	textur	cherri	fruit	light	acid	palat	tannin	soft
8	dri	cherri	herb	tannin	fruit	red	palat	spice	currant	acid
9	structur	fruit	tannin	firm	age	ripe	rich	acid	dens	black
10	black	cherri	pepper	palat	plum	fruit	tannin	dark	spice	blackberri
11	chardonnay	pineappl	butter	toast	vanilla	oak	tropic	acid	rich	fruit
12	cabernet	sauvignon	merlot	blend	franc	petit	verdot	black	tannin	malbec
13	palat	tannin	cherri	berri	black	red	firm	spice	licoric	leather
14	barrel	fruit	oak	spice	age	new	vanilla	cherri	herb	dark
15	appl	pear	green	palat	fresh	acid	citrus	crisp	fruit	light
16	fruit	cherri	red	tannin	vineyard	raspberri	spice	oak	light	palat
17	citrus	peach	white	fruit	acid	blanc	melon	fresh	miner	crisp
18	sweet	cherri	soft	raspberri	candi	simpl	fruit	ripe	oak	blackberri
19	lemon	lime	appl	acid	palat	fresh	dri	crisp	miner	white

Tabela 4.1: Identificação dos *clusters* e seus *tokens* mais significativos.

4.2 RECOMENDAÇÃO DO SISTEMA

Para demonstrar o funcionamento do sistema, foram feitas três solicitações de recomendações, conforme mostrado nas Figuras 4.1 a 4.6. Na interação usuário-aplicação, foram solicitadas recomendações de uvas relacionadas às características desejadas em vinhos e essa informação foi utilizada para calcular o *cluster* mais adequado à solicitação do usuário. Para validar o resultado das indicações e do *cluster* calculado, foi realizada a análise dos *tokens* referentes ao *cluster*, na Tabela 4.1, e às uvas retornadas como recomendações no mapa.

A primeira solicitação foi referente às uvas de vinhos de coloração tinta, conforme mostrado nas Figuras 4.1 e 4.2. O *cluster* calculado para tal foi o de número 12 e, ao analisar a Tabela 4.1, é possível verificar que entre suas expressões de maior significância estão *sauvignon*, *black* e *tannin*, que remetem a características de vinhos tintos. Além disso, as recomendações dadas pelo sistema, nas Figuras 4.1 e 4.2, são de uvas/*blends* majoritariamente conhecidas por serem utilizadas na produção de vinhos tintos [49].

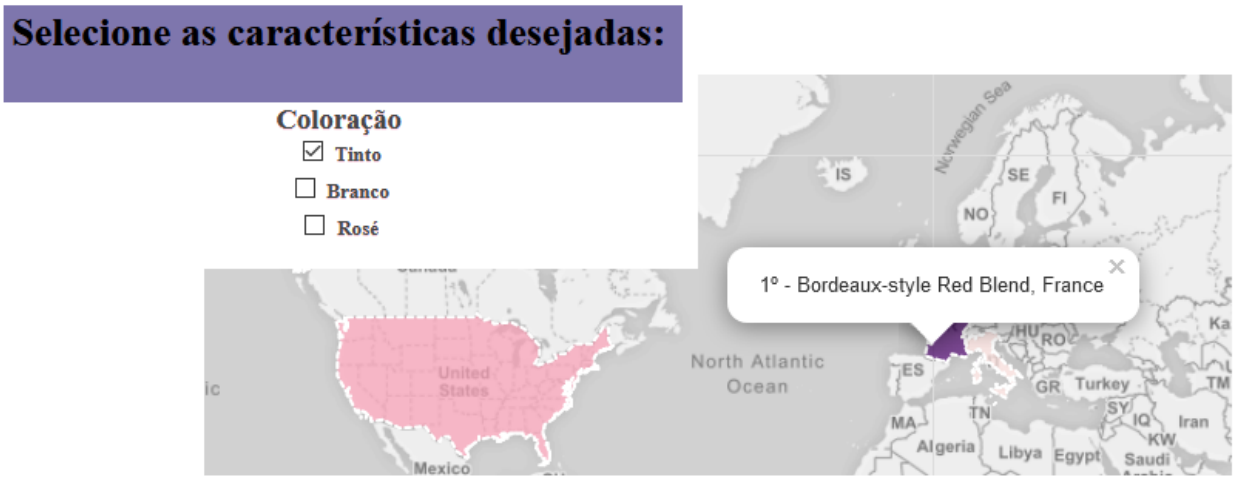

Figura 4.1: Recomendação dada pelo sistema após envio de preferência do usuário por vinho tinto.

Figura 4.2: Recomendação dada pelo sistema após envio de preferência do usuário por vinho tinto.

Na segunda interação com o sistema, foi feita a solicitação de uvas mais correlacionadas com vinhos brancos, suaves e ácidos, que, segundo o sistema, foram melhor descritos pelo *Cluster* 19. Consultando-se novamente a Tabela 4.1, foi possível validar tal informação com *tokens* como *lemon*, *acid* e *white*. E, ao analisar as uvas/*blends* recomendadas (*Sauvignon Blanc*, *White Blend* e *Riesling*) nas Figuras 4.3 e 4.4, chegou-se a conclusão de que o sistema fez recomendações compatíveis, visto as variedades são utilizadas para produção de vinhos brancos e suaves, sendo que estes costumam apresentar acidez [49].

Figura 4.3: Recomendação dada pelo sistema após envio de preferência do usuário por vinho branco, ácido e suave.

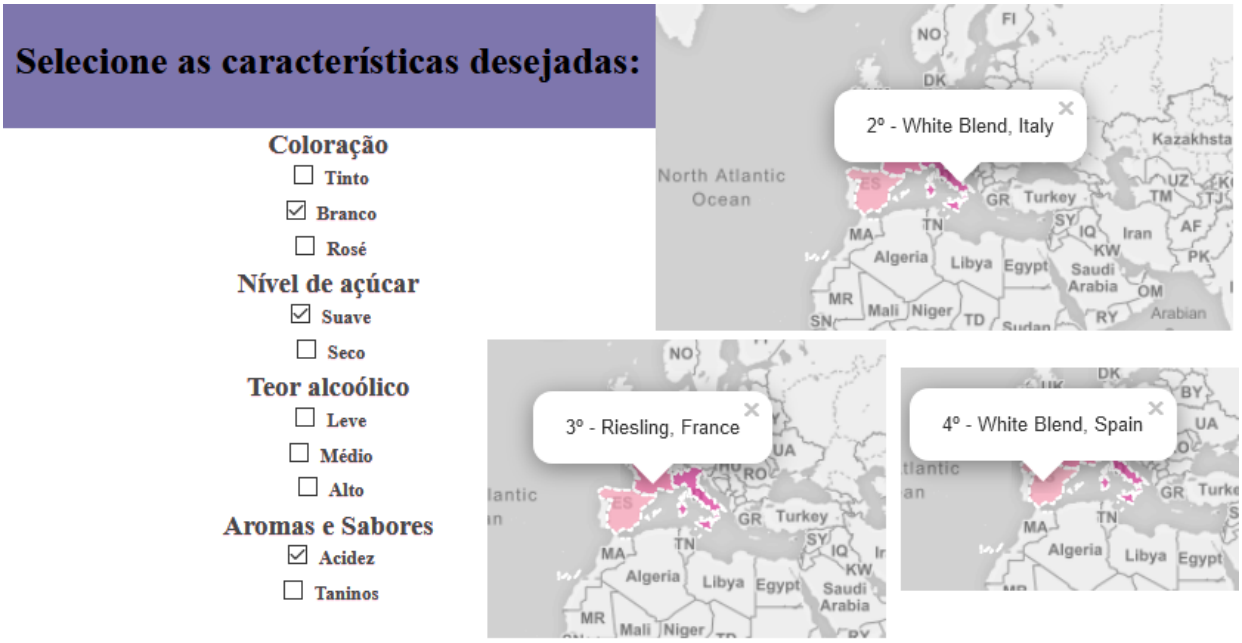

Figura 4.4: Recomendação dada pelo sistema após envio de preferência do usuário por vinho branco, ácido e suave.

A última solicitação de recomendação foi de uvas que tivessem correlação com vinhos *rosé*, frutados e de corpo médio, melhor descritos no *Cluster* 1. As características descritas pelos *tokens* desse *cluster* dizem respeito aos vinhos produzidos a partir de uvas tintas com cor intermediária, notas de frutas vermelhas e ácidos, conforme as Figuras 4.5 e 4.6 .

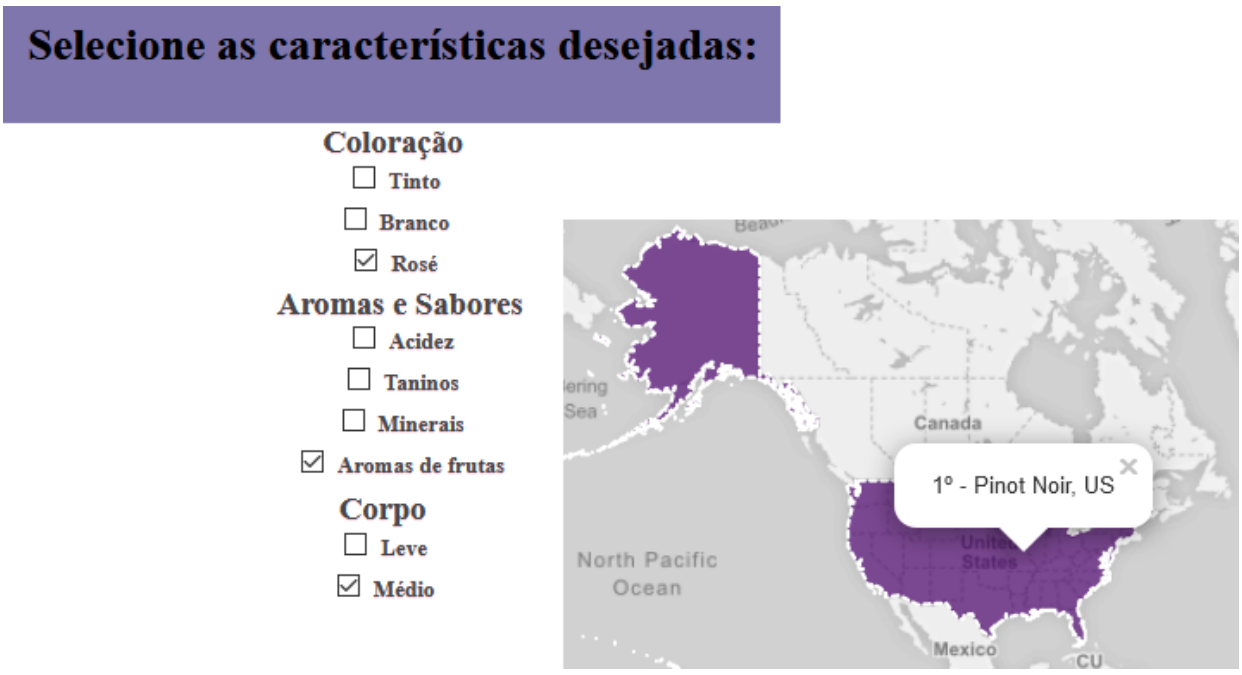

Figura 4.5: Recomendação dada pelo sistema após envio de preferência do usuário por vinho rosé, frutado e de corpo médio.

Figura 4.6: Recomendação dada pelo sistema após envio de preferência do usuário porvinho rosé, frutado e de corpo médio.

É importante ressaltar que o sistema de recomendação proposto não se aplica a todos os vinhos existentes, visto que a análise feita se restringiu a um lote de dados específico. Além disso, como é possível verificar na Figura 4.7, o *dataset* utilizado possuía uma grande quantidade de vinhos com origem nos Estados Unidos, o que acarretou numa maior ocorrência de vinhos estadunidenses em todos os *clusters*, de forma que a métrica de pontuação utilizada para escolher a tupla *(cluster, país, uva, variedade)* favorecia, na maioria dos casos, as tuplas do país. Por isso, foi feita a escolha de enviar as quatro tuplas com a maior pontuação do *cluster*, de modo a não representar tuplas contendo o mesmo país. Nas Figuras 4.8 a 4.14 é possivel verificar a quantidade de vinhos por variedade de uva referente a cada um dos países de origem do *dataset*.

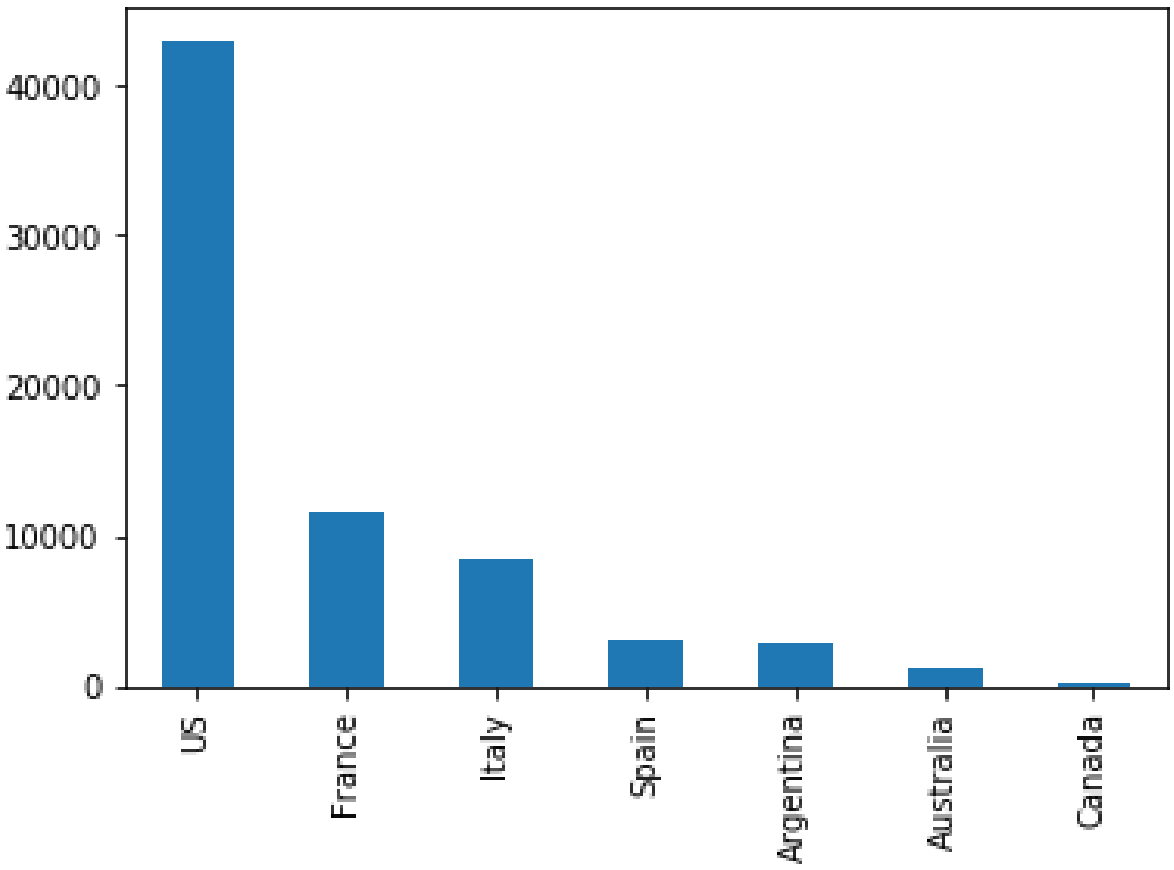

Figura 4.7: Quantidade de vinhos por país no *dataset* pré-processado.

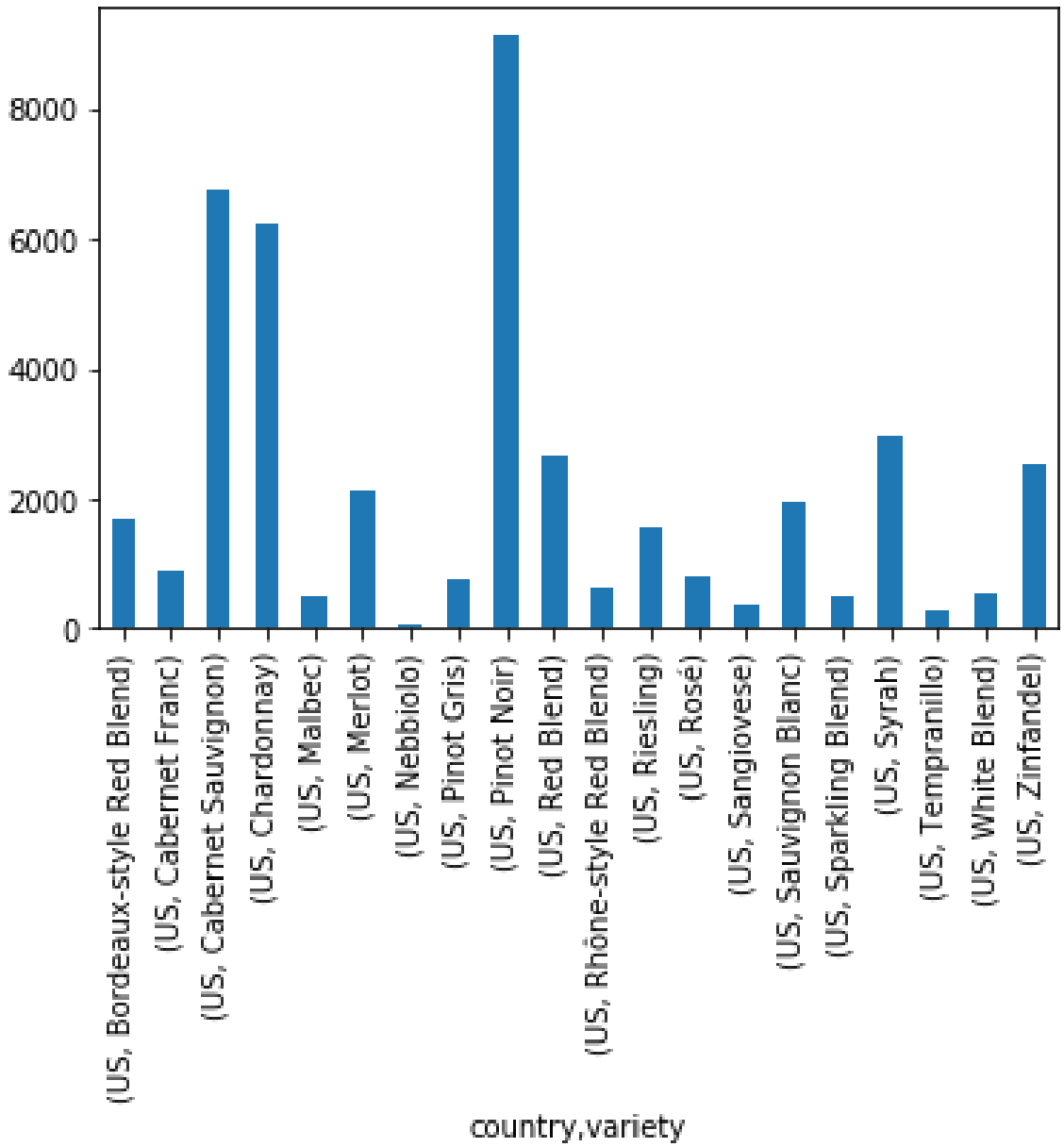

Figura 4.8: Quantidade de vinhos por variedade dos Estados Unidos no *dataset* pré-processado.

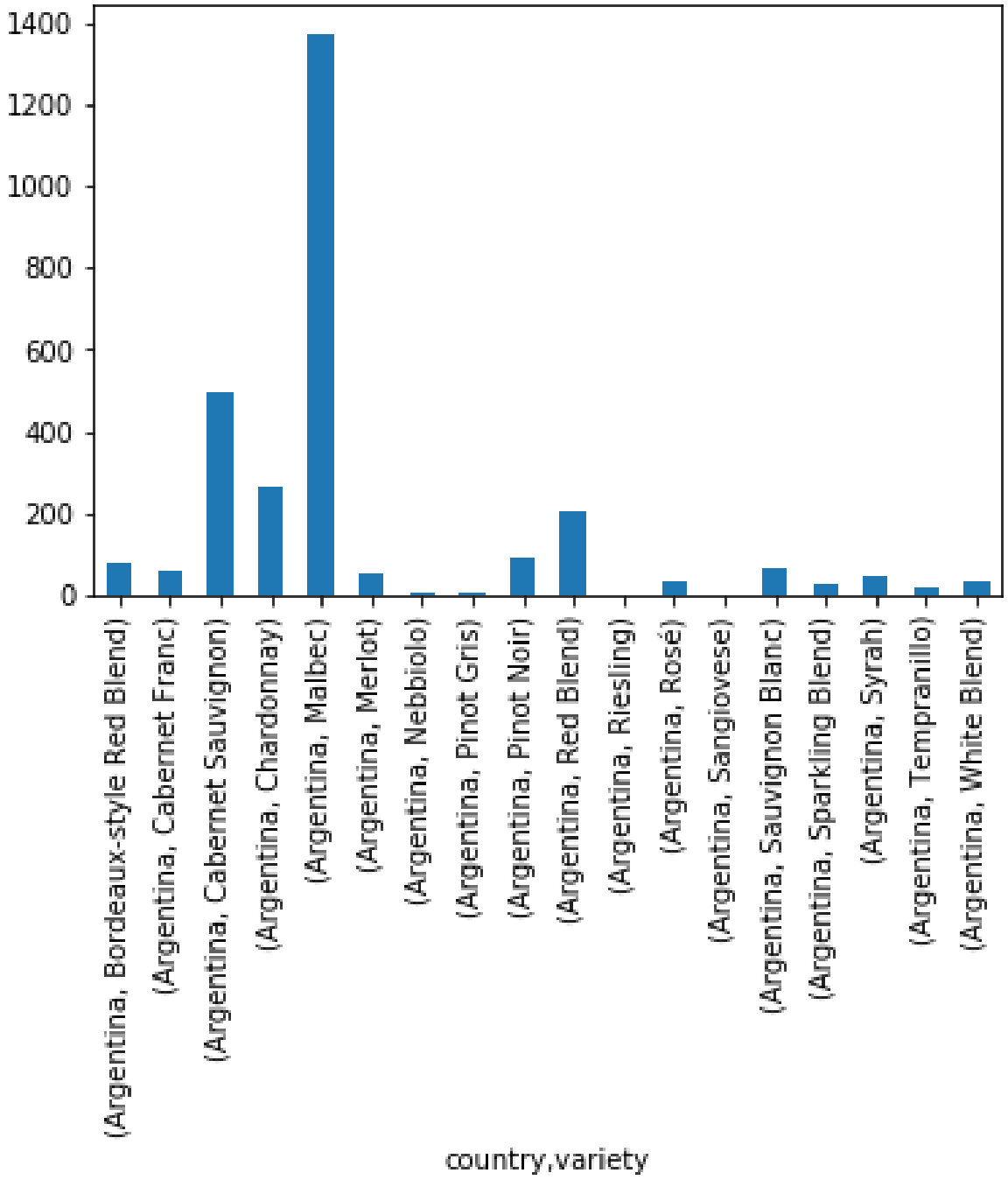

Figura 4.9: Quantidade de vinhos por variedade da Argentina no *dataset* pré-processado.

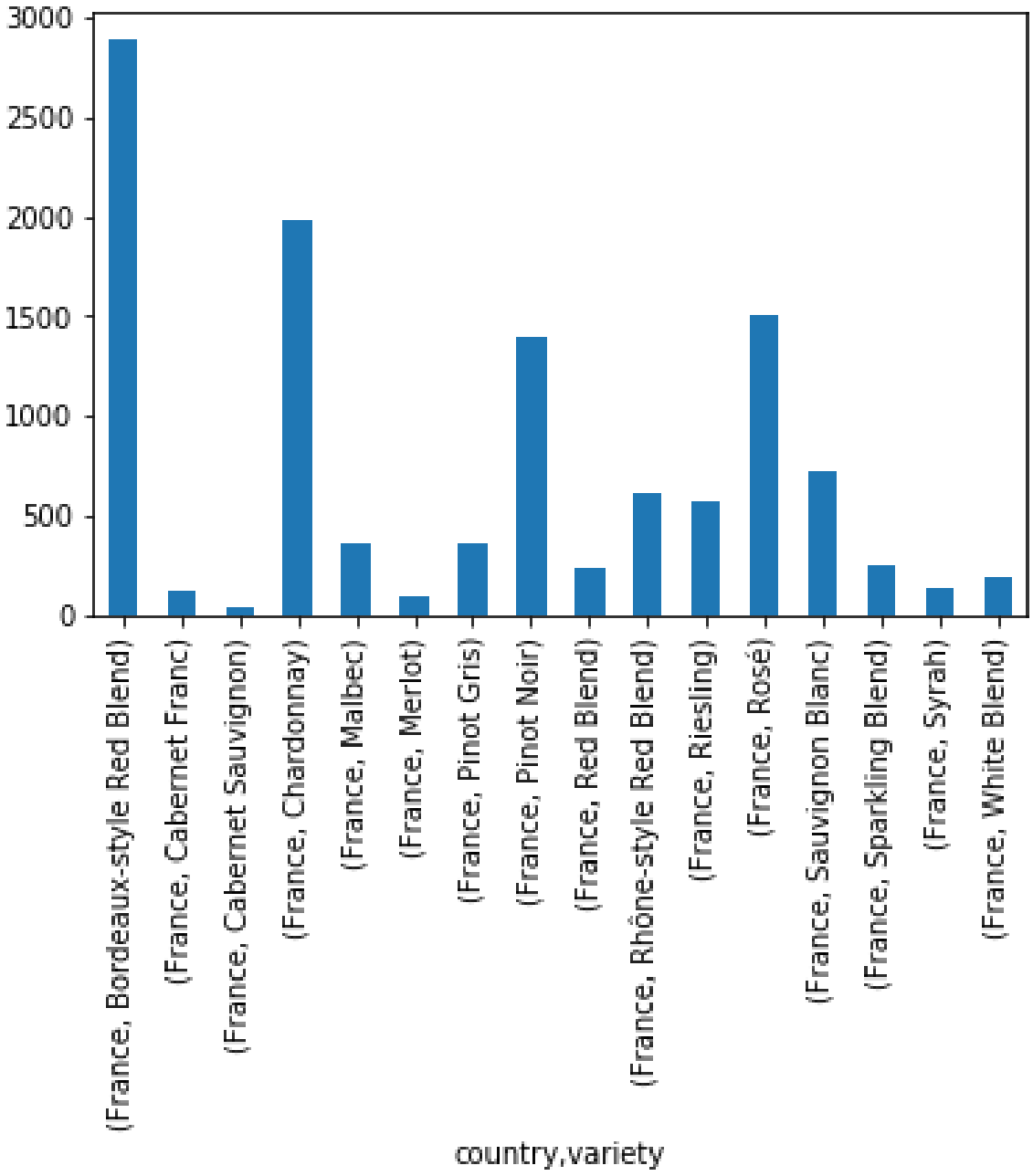

Figura 4.10: Quantidade de vinhos por variedade da França no *dataset* pré-processado.

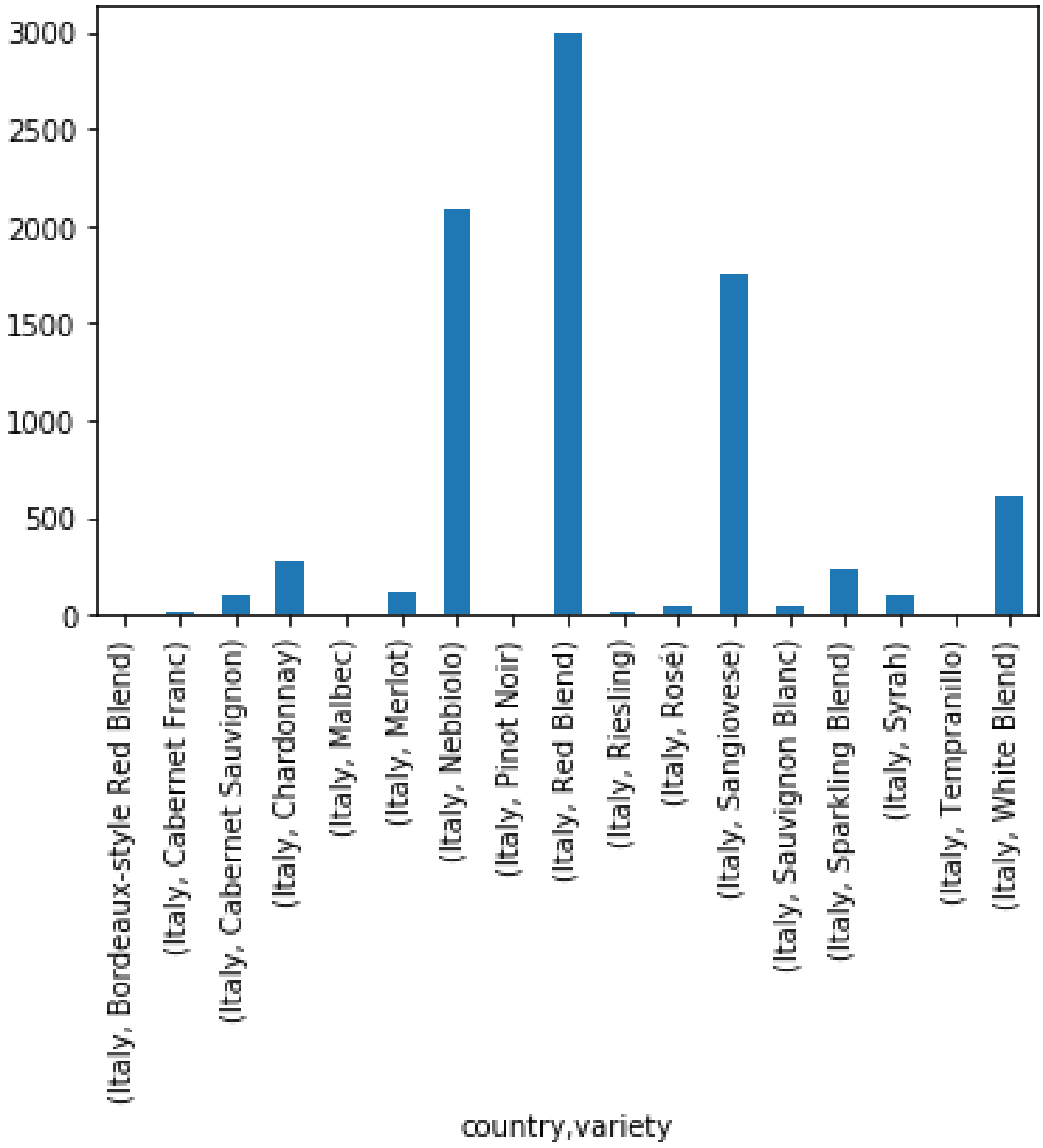

Figura 4.11: Quantidade de vinhos por variedade da Itália no *dataset* pré-processado.

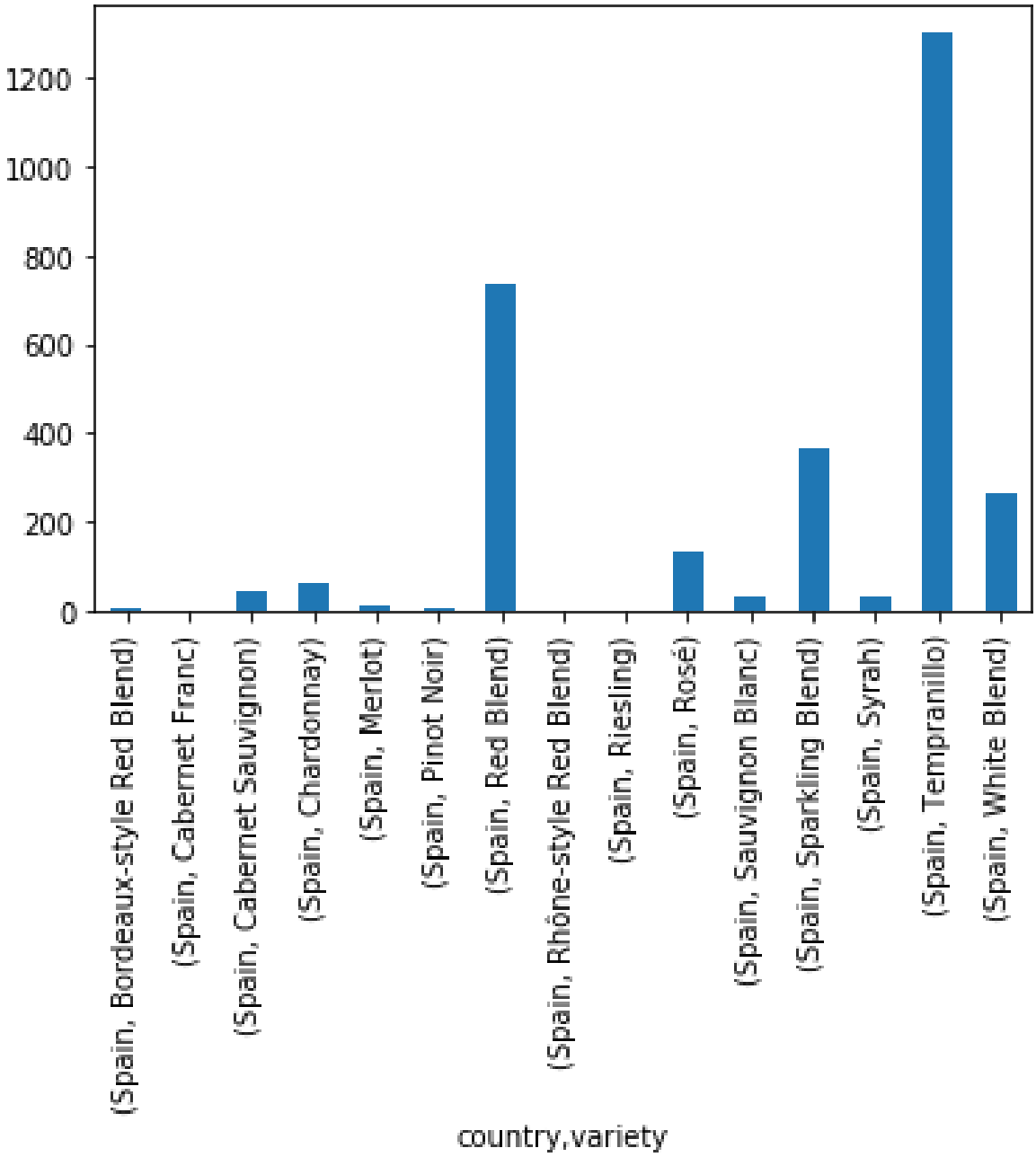

Figura 4.12: Quantidade de vinhos por variedade da Espanha no *dataset* pré-processado.

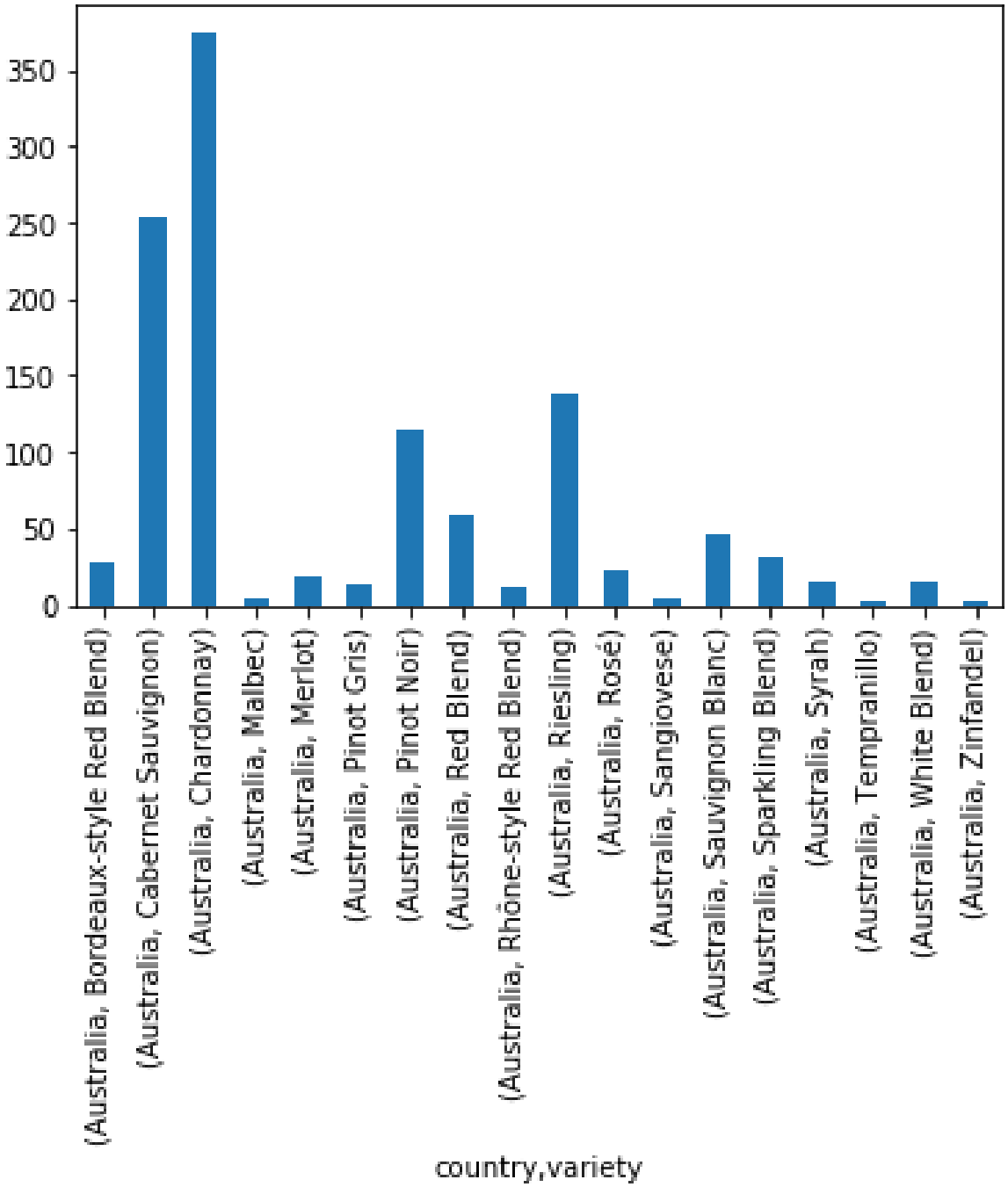

Figura 4.13: Quantidade de vinhos por variedade da Austrália no *dataset* pré-processado.

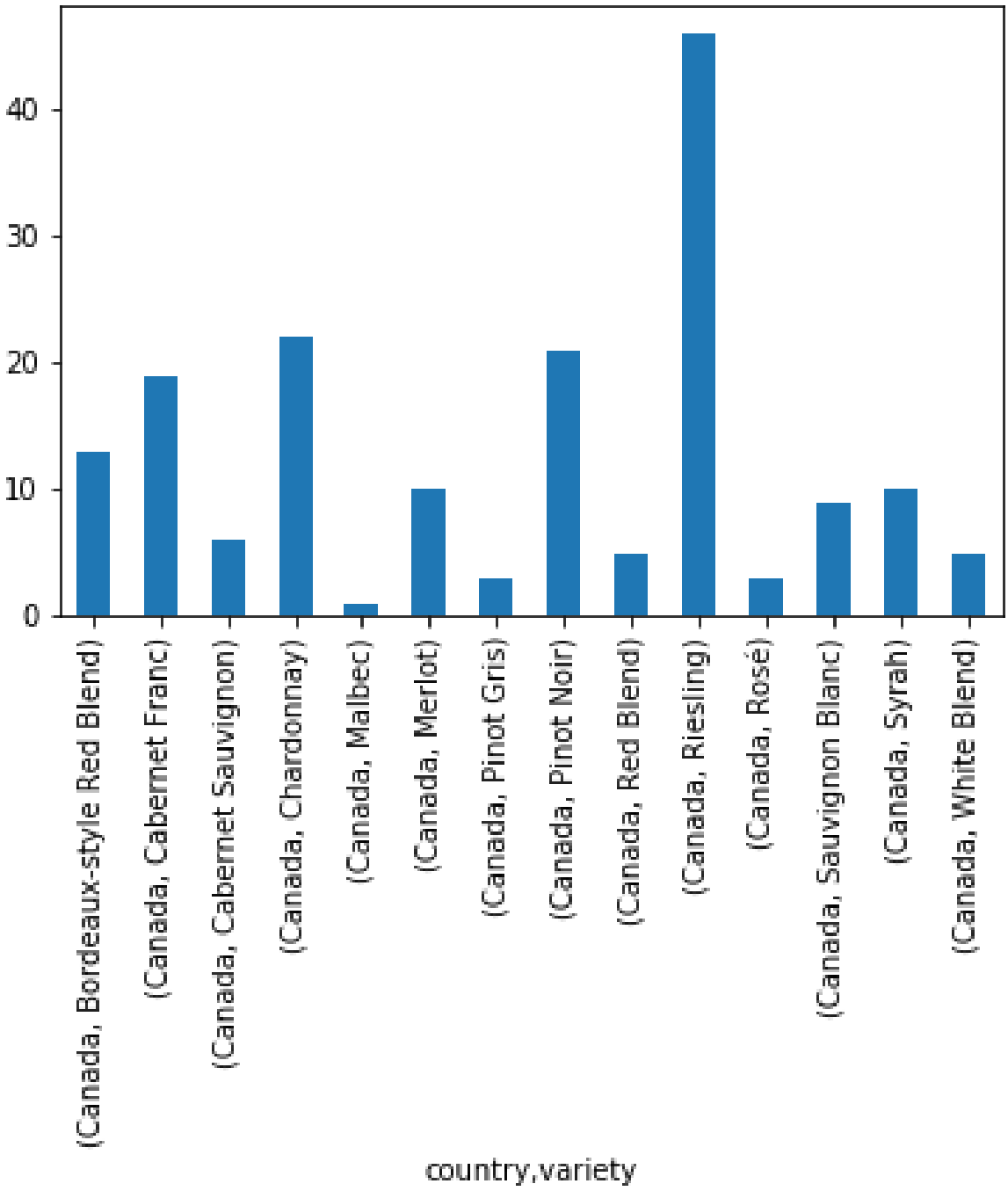

Figura 4.14: Quantidade de vinhos por variedade do Canadá no *dataset* pré-processado.

5 CONCLUSÃO E TRABALHOS FUTUROS

O sistema proposto é capaz de sugerir ao usuário recomendações referentes ao tipo de uva e seu país de origem, de acordo com suas preferências em vinhos, e auxiliar na tomada de decisão. Tal ferramenta pode ser útil tanto para consumidores que desejam fazer escolhas mais adequadas ao realizar uma compra, quanto para empresas que desejam uma ferramenta que auxilie seus clientes no momento da escolha de um produto [4]. Além disso, a lógica do sistema responsável pelo processamento de dados e criação de modelo de agrupamento pode ser aplicada a outros sistemas de recomendação que disponham de descrição e classes de produto.

É possível adicionar funcionalidades no sistema para aumentar a interatividade entre o usuário e aplicação, como:

- plataforma que permita ao usuário escrever as características que deseja, para que o sistema faça a recomendação;
- escalar a aplicação ao nível de aprendizado por reforço, de modo que o usuário possa enviar um *feedback* quanto às recomendações que recebeu, para que o sistema possa recalcular os *clusters* de acordo com o nível de eficiência das recomendações;
- realização de cadastro de vinhos por usuários e automatização dos processamento dos dados, visto que os *clusters* variam conforme o conjunto de dados.

Também podem ser feitos aprimoramentos na forma de sugestão da ferramenta, de forma que a mesma não esteja limitada à recomendação de variedade de uva e seu país, mas contenha, também, indicações de rótulos para o usuário. Informações como pontuação obtida a partir de opiniões de usuários e preço, já adotadas em plataformas de *reviews* de vinhos, podem ser acrescentadas ao modelo de clusterização, de forma que a recomendação também seja feita com base no valor disposto pelo usuário/consumidor.

O sistema, apesar de ter foco no produto da vitivinicultura, com os aprimoramentos citados, também pode ser estendido ao enoturismo, de forma que as recomendações de rótulos venham acompanhadas de sugestões de destinos enoturísticos relacionados a eles.

Por fim, é possível realizar diversas correções na aplicação. Tendo em vista o que foi relatado na Seção 4.2, é necessário acrescentar um nivelador de influência à métrica de pontuação da tupla *(cluster, uva, país)*, para que o sistema entregue dados normalizados ao usuário e não tenda aos itens com maior ocorrência nos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

1 VINE, I. O. of; WINE. *International Organisation of Vine and Wine*. 2019. Disponível em: <http://www.oiv.int/>.

2 ALMEIDA, A. N.; BRAGAGNOLO, C.; CHAGAS, A. A. L. S. A Demanda por Vinho no Brasil: elasticidades no consumo das famÃlias e determinantes da importaÃ§Ã. *Revista de Economia e Sociologia Rural*, scielo, v. 53, p. 433 – 454, 09 2015. ISSN 0103-2003. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-20032015000300433&nrm=iso>.

3 ESTUDO do Mercado Brasileiro de Vinhos Tranquilos e Vinhos Espumantes Quantitativo - Oferta. 2008.

4 SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. E-commerce recommendation applications. *Data mining and knowledge discovery*, Springer, v. 5, n. 1-2, p. 115–153, 2001.

5 CAMPOS, R. de; TAGLIARI, M. *Dicionário do Vinho*. Editora Campos, 2017. ISBN 9788563137449. Disponível em: <https://books.google.com.br/books?id=JphEDgAAQBAJ>.

6 TREVISAN, R. *Moderno Dicionário da Lingua Portuguesa*. [S.l.]: Melhoramentos, 2006. ISBN 8506028604.

7 MCGOVERN STUART J. FLEMING, S. H. K. P. E. *The Origins and Ancient History of Wine*. 1. ed. London: Routledge, 1996. v. 1. (FOOD AND NUTRITION IN HISTORY AND ANTHROPOLOGY (Book 11), v. 1). ISBN 9056995529.

8 GUARINELLO, N. L. A civilização do vinho - um ensaio bibliográfico. *Anais do Museu Paulista: História e Cultura Material*, v. 5, p. 275–278, 1997.

9 LOCKS, E. B.; TONINI, H. Enoturismo: o vinho como produto turístico. *Revista Turismo em Análise*, v. 16, p. 157–173, 2005.

10 LEEUWEN, C. V.; SEGUINL, G. The concept of terroir in viticulture. *Journal of Wine Research*, Routledge, v. 17, n. 1, p. 1–10, 2006. Disponível em: <https://doi.org/10.1080/09571260600633135>.

11 TERSINA. *Rewriting Wine 101: Terroir*. 20162. Disponível em: <https://www.afoodieworld.com/tersina/7592-rewriting-wine-101-terroir>. Acesso em: 08 de Julho de 2019.

12 RESOLUçãO Mercosul Nº 77, DE 01 DE JANEIRO DE 2005. 2005.

13 GUERRA, C. C.; MANDELLI, F.; TONIETTO, J.; ZANUS, M. C.; CAMARGO, U. A. Conhecendo o essencial sobre uvas e vinhos. *Embrapa Uva e Vinho. Documentos*, Bento Gonçalves: Embrapa Uva e Vinho., 2005.

14 J.RUSSEL, P. N. S.; DAVIS, E. *Artificial Intelligence: A Modern Approach.* 3. ed. Upper Saddle River: Prentice Hall, 2010. v. 1. (Prentice Hall series in artificial intelligence., v. 1). ISBN 0136042597.

15 TURING, A. M. Computing machinery and intelligence. *Mind*, v. 1, p. 433–460, 1950.

16 GRANATYR, J. *Teste de Turing*. 2019. Disponível em: <https://iaexpert.com.br/index.php/2016/07/19/historico-da-ia-teste-de-turing/>. Acesso em: 05 de Julho de 2019.

17 QIU, J.; WU, Q.; DING, G.; XU, Y.; FENG, S. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, Nature Publishing Group, v. 2016, n. 1, p. 67, 2016.

18 ERICSSON. *Mobile data traffic report*. 2019. Disponível em: <https://www.ericsson.com/en/mobility-report/mobility-visualizer?f=1&ft=1&r=2,3,4,5,6,7,8,9,13&t=8&s=1,2,3&u=1&y=2018,2024&c=1>. Acesso em: 08 de Julho de 2019.

19 ZHU, X. J. *Semi-supervised learning literature survey*. [S.l.], 2005.

20 SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction.* 2. ed. Cambridge, Massachusetts: Bradford Book, 2014. v. 1. (The MIT PressCambridge., v. 1). ISBN 0262039249.

21 JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v. 31, p. 651–666, 2010.

22 ZHANG, H. *Robot Learning*. 2019. Disponível em: <http://inside.mines.edu/~hzhang/Courses/CSCI473-573/Lectures/07-RobotLearning.pdf>. Acesso em: 08 de Julho de 2019.

23 C.DUBES, A. K. J. R. *Algorithms for Clustering Data*. 1. ed. Engle Cliffs, New Jersey 07632: Prentice Hall, 1988. v. 1. (Prentice Hall series in artificial intelligence., v. 1). ISBN 013022278X.

24 LADEIRA, A. P. *Processamento de linguagem natural : caracterização da produção científica dos pesquisadores brasileiros*. Tese (Doutorado) — Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação, 11 2010.

25 HOTHO, A.; NURNBERGER, A.; PAASS, G. A brief survey of text mining. In: CITESEER. *Ldv Forum*. [S.l.], 2005. v. 20, n. 1, p. 19–62.

26 DALE, R. Classical approaches to natural language processing. In: *Handbook of natural language processing, second edition*. [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7.

27 PALMER, D. D. Text preprocessing. In: *Handbook of natural language processing, second edition*. [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7.

28 HIPPISLEY, A. Lexical analysis. In: *Handbook of natural language processing, second edition*. [S.l.]: CRC Press, Taylor & Francis Group, 2010. p. 3–7.

29 OSA, M. Garma de la; SáNCHEZ, Y. Map overlay problem. *Handbook of Research on Geoinformatics*, p. 65–72, 01 2009.

30 ALBA CARLOS ALBERTO FLORES, A. M. L. M. V. J. M. F. Sig para a gestão vitivinícola no vale dos vinhedos, rs. *Embrapa Clima Temperado - Capítulo em livro científico (ALICE)*, In: BERNARDI, AC de C.; NAIME, J. de M.; RESENDE, AV de; BASSOI, LH; INAMASU . . . , 2014.

31 MATHEWS, A. J. Applying geospatial tools and techniques to viticulture. *Geography Compass*, v. 7, p. 22–34, 2013.

32 FOWLER, M. *Patterns of enterprise application architecture*. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 2002.

33 ROSSUM, G. Python reference manual. CWI (Centre for Mathematics and Computer Science), 1995.

34 OLIPHANT, T. E. Python for scientific computing. *Computing in Science & Engineering*, IEEE, v. 9, n. 3, p. 10–20, 2007.

35 PYTHON. *The Python Tutorial*. 2016. Disponível em: <https://docs.python.org/3/tutorial/index.html>. Acesso em: 05 de Junho de 2016.

36 AYER, V. M.; MIGUEZ, S.; TOBY, B. H. Why scientists should learn to program in python. *Powder Diffraction*, Cambridge University Press, v. 29, n. S2, p. S48–S64, 2014.

37 MCKINNEY, W. pandas: a foundational python library for data analysis and statistics.

38 LOPER, E.; BIRD, S. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

39 OLIPHANT, T. E. *A guide to NumPy*. [S.l.]: Trelgol Publishing USA, 2006. v. 1.

40 BAYER, M. Sqlalchemy-the database toolkit for python. *URL http://www. sqlalchemy. org/. Accessed on the 13th of November*, 2012.

41 PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

42 ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database Systems*. [S.l.]: Pearson, 2010. v. 1. ISBN 0136086209.

43 ORACLE. *Database Documentation*. 2019. Disponível em: <https://docs.oracle.com/en/database/>. Acesso em: 08 de Julho de 2019.

44 ZONE, M. D. *MySQL 5.7 Reference Manual*. [S.l.], 2019. Disponível em: <https://downloads.mysql.com/docs/refman-5.7-en.pdf>. Acesso em: 08 de Julho de 2019.

45 WINE Reviews. 2019. Disponível em: <hhttps://www.kaggle.com/zynicide/wine-reviews>. Acesso em: 20 de Fevereiro de 2019.

46 THOUTT, Z. *Wine Deep Learning*. 2019. Disponível em: <https://github.com/zackthoutt/wine-deep-learning>. Acesso em: 20 de Fevereiro de 2019.

47 KITAGAWA, K. *Exploring Wine Descriptions with NLP and kMeans*. 2019. Disponível em: <https://github.com/kitakoj18/wine_descp>. Acesso em: 15 de Abril de 2019.

48 TF-IDF term weighting. 2019. Disponível em: <https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting>.

49 WINE Frog: Your wine education resource on the web. 2019. Disponível em: <https://www.winefrog.com>. Acesso em: 16 de Junho de 2019.

50 LEAFLET. *Leaflet documentation*. 2019. Disponível em: <https://leafletjs.com/reference-1.5.0.html>. Acesso em: 16 de Junho de 2019.