**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: | | **Data início curso**: \_\_/\_\_/\_\_\_\_ |
| **Orientador**: | | **Defesa em:**  Mês/Ano |
| **Curso**: Selecionar. | **Modalidade**: Selecionar. | Turma: Selecionar. |

1. **Título do projeto** *(Inicial)*

**Previsão de Avaliação Geral de Cervejas com Base em Características Sensoriais Usando Técnicas de Machine Learning**

1. **Introdução**

*(*Nesta seção o autor deve contextualizar o assunto, estabelecendo uma relação com outros trabalhos publicados sobre o tema. Da mesma maneira, deve ser apresentada a justificativa para a realização do trabalho, situando a importância do problema estudado e apoiada em citações bibliográficas de fontes confiáveis. *O texto deverá ser escrito de forma impessoal.,* redigido em no máximo duas páginas e não deve conter subtópicos, figuras ou tabelas)

O mercado de cervejas artesanais tem crescido exponencialmente ao longo dos últimos anos, levando a um aumento na diversidade de estilos, sabores e características sensoriais disponíveis para os consumidores. Ao mesmo tempo, plataformas de avaliação de cervejas, como aplicativos e sites especializados, permitem que milhões de usuários compartilhem suas impressões e notas sobre diversos aspectos das bebidas que consomem. Dentre os critérios mais comumente avaliados, encontram-se o aroma, a aparência, o paladar e o gosto, além de uma avaliação geral que sintetiza a experiência do consumidor.

Este aumento de dados disponibilizados por consumidores representa uma oportunidade valiosa para aplicar técnicas de ciência de dados com o objetivo de entender e prever como diferentes características de uma cerveja podem influenciar a avaliação geral que ela receberá. O uso de modelos preditivos, como regressão linear, árvores de decisão, e algoritmos de ensemble como Random Forest e XGBoost, oferece uma abordagem robusta para investigar esse fenômeno.

Estudos anteriores sobre previsão de preferências de consumidores já demonstraram a eficácia de técnicas de machine learning em domínios como recomendação de filmes, previsão de sucesso de músicas e até mesmo avaliações de vinhos. No entanto, a previsão da avaliação geral de cervejas com base em suas características sensoriais ainda é um campo pouco explorado, especialmente considerando a rica diversidade de variáveis envolvidas nesse processo de avaliação.

Este projeto busca preencher essa lacuna ao explorar um dataset contendo mais de 1,5 milhões de avaliações de cervejas realizadas por consumidores. As avaliações abrangem múltiplos critérios sensoriais, como aroma, aparência, paladar e gosto, e permitem a análise de como essas variáveis podem prever a nota geral atribuída pelos consumidores.

A relevância deste estudo está ancorada em duas frentes principais. Primeiro, do ponto de vista acadêmico, ele contribui para o campo de previsão de preferências de consumidores, com foco em bebidas, um setor que apresenta variações complexas de qualidade percebida. Segundo, do ponto de vista prático, os resultados podem beneficiar diretamente cervejarias e distribuidores, que poderiam ajustar suas produções com base nas características mais valorizadas pelos consumidores, aumentando suas chances de sucesso no mercado.

1. **Objetivo**

*(Qual o objetivo principal do trabalho, ou seja, qual pergunta deve ser respondida ao final da sua pesquisa* *)*

O objetivo principal deste trabalho é desenvolver e avaliar modelos de machine learning capazes de prever a avaliação geral de uma cerveja com base em suas características sensoriais (aroma, aparência, paladar e gosto) e no teor alcoólico. Ao final da pesquisa, espera-se responder à seguinte pergunta: Quais características sensoriais têm maior impacto na avaliação geral de uma cerveja e como essas características podem ser usadas para prever a nota atribuída pelos consumidores?

1. **Material e Métodos**

*(Descrever o(s) método(s) de coleta de dados e a(s) ferramenta(s) de análise a ser(em) utilizada(s) no trabalho de conclusão de curso, ou seja, como será a condução da pesquisa e a forma de obtenção dos resultados, por exemplo, fontes de dados, técnicas, procedimentos, índices, entre outros)*

Fonte de Dados

A base de dados utilizada neste estudo será composta por um dataset contendo mais de 1,5 milhões de avaliações de cervejas. Este dataset inclui as seguintes colunas: brewery\_name, review\_time, review\_overall, review\_aroma, review\_appearance, review\_profilename, beer\_style, review\_palate, review\_taste, beer\_name e beer\_abv (álcool por volume).

Pré-processamento

O primeiro passo será o pré-processamento dos dados, incluindo a limpeza e a preparação dos mesmos para a análise. Serão aplicados os seguintes procedimentos:

Tratamento de dados faltantes: Serão identificadas e tratadas entradas com dados ausentes.

Conversão de formatos: A coluna de tempo (review\_time) será convertida em um formato apropriado (timestamp) para análises temporais, caso necessárias.

Normalização de variáveis: As variáveis numéricas (como avaliações de aroma, aparência, etc.) serão normalizadas para facilitar o treinamento dos modelos.

Criação de variáveis categóricas: Variáveis como beer\_style e brewery\_name serão transformadas em categorias numéricas utilizando técnicas como One-Hot Encoding.

Modelos Preditivos

O próximo passo será a criação e a avaliação de diferentes modelos de machine learning. Os modelos serão treinados utilizando a variável alvo review\_overall e as variáveis preditoras serão as avaliações de aroma, aparência, paladar, gosto e o teor alcoólico. Os seguintes modelos serão aplicados:

Regressão Linear: Modelo inicial para estabelecer uma baseline de comparação.

Árvores de Decisão: Para capturar possíveis interações não-lineares entre as variáveis.

Random Forest: Algoritmo de ensemble que utiliza múltiplas árvores de decisão para aumentar a precisão preditiva.

XGBoost: Um dos modelos mais avançados para tarefas de regressão, que otimiza o desempenho através de boosting.

Métricas de Avaliação

Os modelos serão avaliados utilizando as seguintes métricas:

R² (Coeficiente de Determinação): Mede a proporção da variabilidade total explicada pelo modelo.

MAE (Erro Absoluto Médio): Mede o erro médio entre os valores previstos e os valores reais.

RMSE (Raiz do Erro Quadrático Médio): Avalia o quão grande é o erro médio das previsões.

Validação e Teste

Para garantir a robustez dos resultados, será utilizado um processo de validação cruzada k-fold com 5 dobras, onde o dataset será dividido em partes e cada parte será usada para treinar e testar os modelos. Isso garantirá que os resultados não sejam enviesados por um subconjunto específico de dados.

1. **Resultados Esperados**

*(Descrever os resultados que são esperados após a realização da coleta e análise dos dados, ou seja, quais resultados são esperados ao final da pesquisa)*

Espera-se que, ao final da pesquisa, os modelos de machine learning treinados sejam capazes de prever a avaliação geral de uma cerveja com base em suas características sensoriais e teor alcoólico com um nível significativo de precisão. Entre os resultados esperados estão:

Identificação dos fatores mais importantes: A partir da análise de importância de variáveis, espera-se identificar quais características sensoriais (aroma, aparência, paladar ou gosto) têm o maior impacto na avaliação geral.

Modelo preditivo eficiente: Um modelo preditivo com bom desempenho, medido pelas métricas R², MAE e RMSE, que poderá ser utilizado para prever a nota de uma cerveja com base em suas características.

Contribuição para a indústria: Os resultados poderão fornecer insights úteis para cervejarias ajustarem suas fórmulas ou focarem em características que são mais valorizadas pelos consumidores, aumentando sua competitividade no mercado.

A pesquisa também poderá revelar possíveis padrões interessantes nas avaliações, como a existência de sesgos em relação a certos estilos de cerveja ou fabricantes, o que pode gerar novas perguntas de pesquisa ou linhas de investigação futura.

1. **Cronograma de Atividades**

*(Adicionar as “Atividades planejadas”, assim como o período planejado (tempo para desenvolver cada atividade) para a realização de cada atividade, sendo que deverá ser adequado ao calendário de entregas das etapas do trabalho de conclusão de curso definido pela coordenação. Marcar com um “x” a coluna que corresponde ao período planejado para desenvolver cada atividade)*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Definição do tema e levantamento bibliográfico | **X** | X |  |  |  |  |  |  |  |  |
| Pré-processamento e limpeza dos dados |  | **X** | **X** |  |  |  |  |  |  |  |
| Análise exploratória dos dados (EDA) |  | **X** | **X** |  |  |  |  |  |  |  |
| Seleção e ajuste dos modelos preditivos |  |  | **X** | **X** |  |  |  |  |  |  |
| Treinamento dos modelos |  |  | **X** | **X** |  |  |  |  |  |  |
| Avaliação dos modelos (validação cruzada) |  |  |  | **X** |  |  |  |  |  |  |
| Otimização e ajuste fino dos modelos |  |  |  | **X** |  |  |  |  |  |  |
| Redação preliminar do relatório |  |  |  | **X** | **X** |  |  |  |  |  |
| Revisão do relatório e ajustes |  |  |  | **X** | **X** |  |  |  |  |  |
| Apresentação e defesa do TCC |  |  |  |  | **X** |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

*(Listagem das bibliografias citadas no projeto de pesquisa, seguindo rigorosamente as Normas do MBA USP/Esalq. Consulte o manual de “Normas para Elaboração do Trabalho de Conclusão de Curso” disponível no Sistema TCC)*