UNIVERSIDADE FEDERAL DE CATALÃO
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E OTIMIZAÇÃO

WELISON DE CAMARGO VIEIRA

ANALISANDO CERVEJA ARTESANAL POR MEIO DE 3 MODELOS DE CLASSIFICAÇÃO E APRENDIZADO DE MÁQUINA.

CATALÃO – GO

2023

UNIVERSIDADE FEDERAL DE CATALÃO
INSTITUTO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E OTIMIZAÇÃO

WELISON DE CAMARGO VIEIRA

ANALISANDO CERVEJA ARTESANAL POR MEIO DE 3 MODELOS DE CLASSIFICAÇÃO E APRENDIZADO DE MÁQUINA.

Trabalho apresentado ao PPGMO como requisito parcial para obtenção dos créditos da disciplina de Inteligência Artificial.

CATALÃO – GO

2023

SUMMARY

1 – INTRODUÇÃO

A cerveja é uma das bebidas mais apreciadas do mundo, sendo também a bebida alcoólica mais consumida no Brasil, país que se destaca como seu terceiro maior produtor. De acordo com a Associação Brasileira da Indústria da Cerveja, isso representa um impacto de 1,6% do PIB brasileiro, movimentando aproximadamente R$ 21 Bilhões de reais apenas em impostos (CERVBRASIL, 2018).

Neste cenário, vemos uma prática conhecida pelo consumidor, ou seja, a cerveja artesanal. Embora não sejam tão representativas se comparadas ao restante do setor em termos econômicos ou mesmo em quesitos de popularidade, as empresas e/ou pessoas responsáveis por essa produção detém uma margem crescente de consumo, aumentando ainda mais a popularidade da bebida e mudando ligeiramente a cultura popular sobre o assunto.

Este fenômeno parece estar relacionado a diferenças claras de sabor e variedade que a produção artesanal permite. Segundo DIAS (2018), a mídia televisiva brasileira costuma apenas apoiar marcas líderes e já renomadas de cerveja. Isso, somado ao fato de que os grandes grupos concorrentes do mercado brasileiro de bebidas passaram a comprar indústrias de cerveja artesanal, evidenciam a força da bebida artesanalmente preparada.

Este trabalho analisou de forma exploratória um banco de dados de 73 mil receitas de cerveja artesanal encontrado no domínio "*brewersfriend.com*" e disponibilizado na plataforma "*Kaggle*". O objetivo dessa pesquisa foi a aplicação de métodos de aprendizado de máquina visando a correta classificação dos dados disponíveis nas receitas, além de permitir que através desses resultados seja possível realizar previsões mais assertivas de produto final, possibilitando orientar certas decisões durante o processo de fabricação.

Devemos mencionar que algumas das análises realizadas, referenciam métodos compartilhados em domínios públicos, ainda que comumente aplicados por pesquisadores que abertamente assumem não conhecer o processo de fabricação. O impacto do distanciamento entre pesquisadores e cervejeiros[1], resulta em diversas confusões a respeito da necessidade de certas etapas e sua influência no produto final.

Por fim, ao utilizarmos 3 modelos para diferenciação entre os estilos de cerveja, sendo eles o *"Random Forest Classifier"*, a "*Logistic Regression"* e o "*XGBoost Classifier"*. Cada modelo apresentou resultados diferentes de precisão, sendo necessário definir parâmetros adicionais de análise e compensação. Tais parâmetros, somados a atualização incremental do banco de dados, permitirá a modelagem de propostas viáveis de pesquisas futuras.

[1] Cervejeiro/Mestre Cervejeiro: Aquele que faz cerveja; consultado em "meudicionario.org" e "dicio.com.br".

2 – REVISÃO BIBLIOGRÁFICA

Antes de mais nada, devemos ressaltar o fato de que hoje em dia, a fabricação de cerveja mantém seu próprio espaço entre ser considerada por muitos uma ciência, um tipo de arte, ou mesmo ambos, expandindo em diversos aspectos. Isso também significa que cada parte do processo poderia facilmente render um estudo de caso independente, o que leva a necessidade de orientarmos nossa pesquisa apenas aos métodos de classificação utilizados e seus resultados.

Por sua vez, no contexto do aprendizado de máquina, a classificação, é uma tarefa fundamental que visa atribuir rótulos ou categorias a dados com base em suas características. Entre os algoritmos amplamente adotados para esse fim, selecionamos o *"Random Forest Classifier"*, a "*Logistic Regression"* e o "*XGBoost Classifier"*.

O método "*Random Forest Classifier"* emprega uma abordagem baseada em árvores de decisão. Durante o processo de treinamento, várias árvores são construídas, enquanto suas previsões são combinadas para obter a classificação final. Cada árvore é treinada mediante a seleção de um subconjunto aleatório de características aplicada a um subconjunto único dos dados. De acordo com Liaw e Wiener (2002), esse algoritmo cria uma "coleção de árvores de decisão que são combinadas para obter uma classificação mais precisa e estável". Isso resulta em um modelo robusto, com o intuito de evitar o "*overfitting*[2]". Além de áreas tradicionais, como detecção de fraudes e classificação de imagens, o *"Random Forest Classifier"* tem sido aplicado a estatísticas de várias bases de dados para prever características sensoriais e de sabor com base em ingredientes e processos de fabricação de receitas de cerveja, vinho, uísque, entre outros (Trappey et al., 2016).

O método de *Logistic Regression*, apesar de seu nome, é uma técnica linear usada principalmente para problemas de classificação binária. Ele modela a relação entre características observadas e a probabilidade de pertencer a uma classe específica usando a função logística. Este método é particularmente eficaz quando se trata de interpretabilidade e simplicidade. Conforme apontado por Hosmer e Lemeshow (2004), o "*Logistic Regression"* avalia "a relação entre uma variável dependente binária e uma ou mais variáveis independentes". Em termos de aplicação a estatísticas de bancos de dados de receitas de cerveja, é possível usar a "*Logistic Regression"* para prever se uma cerveja específica pertence a uma categoria de sabor específica com base em seus ingredientes e parâmetros de fabricação.

[2] Um overfitting ocorre quando, nos dados de treino, o modelo tem um desempenho excelente, porém quando utilizamos os dados de teste o resultado é ruim. <https://didatica.tech/underfitting-e-overfitting>

O "*XGBoost Classifier"* é um algoritmo que se destacou por seu desempenho em competições de aprendizado de máquina. Ele constrói uma sequência de árvores de decisão, permitindo cada nova árvore buscar e corrigir os erros das árvores anteriores. O algoritmo otimiza uma função objetivo combinando perda do modelo com termos de regularização. De acordo com Chen e Guestrin (2016), o XGBoost é "uma ferramenta eficaz e escalável para aprender modelos de aumento". Além das funções de classificação utilizadas nesse trabalho, o "*XGBoost Classifier"* poderia também ser usado para prever a popularidade de uma cerveja com base em suas características, permitindo que os cervejeiros ajustem as receitas de acordo com as preferências dos consumidores.

Ambos os modelos "*Random Forest Classifier"*, a "*Logistic Regression"* e o "*XGBoost Classifier"* são abordagens consideravelmente eficientes para problemas de classificação no aprendizado de máquina, cada um com seus atributos e aplicações específicas. Sua escolha, dependerá das características do problema em questão, bem como do tamanho/natureza dos dados, além é claro dos recursos computacionais disponíveis.

3 – METODOLOGIA

Neste estudo, realizamos uma análise detalhada das características de várias receitas de cerveja, aplicando três algoritmos populares de classificação: *"Random Forest Classifier"*, a "*Logistic Regression"* e o "*XGBoost Classifier".*

A princípio, foram reunidas diversas informações detalhadas de bancos de dados disponíveis através da plataforma *"Kaggle"*. Logo em seguida, fizemos uma análise do método de inserção de receitas do site *"brewersfriend.com"*, responsável pelo banco de dados que obtivemos. Ao compreendermos o funcionamento dos campos relacionados aos ingredientes utilizados, métodos de fabricação empregados e informações sensoriais das cervejas, observamos a necessidade de padronização de caracteres.

Utilizando a linguagem Python para construção do código e após a importação, realizamos uma etapa de tratamento utilizando a biblioteca "*SimpleImputer*" para lidar com valores ausentes. Essa ferramenta preencheu lacunas nos dados, usando meios ou valores mais frequentes, garantindo a integridade dos conjuntos de dados (Hastie et al., 2009).

Em seguida, a biblioteca *"StandardScaler"* foi aplicada para padronizar as características numéricas. Essa padronização é fundamental, especialmente para algoritmos como *"Random Forest"* e *"XGBoost"*, que podem ser sensíveis a diferenças de escala entre variáveis (James et al., 2013).

Para um tratamento profundado dos dados, recorremos às bibliotecas *"numpy"*, *"pandas"*, *"matplotlib.pyplot"* e *"seaborn"* para análise exploratória. Usando essas ferramentas, calculamos estatísticas descritivas e criamos visualizações gráficas para avaliar a distribuição das características e identificar padrões.

Para garantir a qualidade dos resultados, identificamos os valores atípicos que poderiam influenciar indevidamente os modelos de classificação (Tukey, 1977), resultando no próximo passo, que foi a divisão dos dados em conjuntos de treinamento e teste, essencial para avaliar a capacidade dos modelos de generalizar para novos dados.

Com os dados devidamente preparados, implementamos os modelos de "*Random Forest Classifier"*, a "*Logistic Regression"* e o "*XGBoost Classifier"* usando as bibliotecas *"scikit-learn"* e *"xgboost"*. Os modelos foram treinados nos dados tratados na etapa anteriormente mencionada, permitindo-lhes aprender os padrões subjacentes nas receitas de cerveja.

Após o treinamento, avaliamos o desempenho de cada modelo usando métricas de análise da biblioteca "accuracy_score". Essas métricas nos forneceram insights sobre a capacidade de cada modelo de fazer previsões precisas.

Comparando os resultados dos três modelos, fomos capazes de determinar qual teve o melhor desempenho na análise das receitas de cerveja. Essa comparação nos permitiu identificar quais algoritmos são mais adequados no contexto aplicado, para prever características específicas das cervejas com base em suas receitas e avaliações sensoriais.

Ao longo do processo, as bibliotecas *"numpy"* e *"pandas"* desempenharam um papel crucial no tratamento dos dados e na realização de cálculos estatísticos. A biblioteca *"matplotlib.pyplot"* foi fundamental para criar gráficos, enquanto o *"seaborn"* adicionou uma camada extra de visualização e interpretação dos resultados.

Assim, por meio dessas etapas, este trabalho descreve desde a coleta e pré-processamento dos dados até a implementação e avaliação dos modelos de classificação. Ao combinar diferentes bibliotecas Python e algoritmos de classificação, fomos capazes de realizar uma análise abrangente das receitas encontradas no banco de dados selecionado, identificando padrões e comparando o desempenho dos modelos utilizados.

4 – EXECUÇÃO, RESULTADOS E DISCUSSÃO

4.1 Dados

Os dados encontrados no Dataset utilizado, incluem um total de 73861 receitas. Apesar de serem relativamente abrangentes, essas receitas representam uma quantidade consideravelmente pequena de dados, o que conta a favor do custo computacional dos algoritmos e possibilita a execução dos códigos em ambientes virtuais de processamento, como é o caso da plataforma "*Colab*[3]" da Google.

Ao baixar o arquivo compactado que possuía os dados do Dataset, reorganizamos os dados, enviando logo em seguida para um sistema de hospedagem próprio, sendo este importado para o ambiente de desenvolvimento. Note na imagem a seguir que também utilizamos alguns parâmetros para garantir a codificação de texto e evitar certos caracteres inválidos.

```
beer_recipe = pd.read_csv('                    .com/tp/recipeData.csv', index_col='BeerID', encoding='latin1')
beer_recipe.head()
```

BeerID	Name	URL	Style	StyleID	Size(L)	OG	FG	ABV	IBU	Color	...	BoilTime	BoilGravity	Efficiency	MashThickness	SugarScale	BrewMethod
1	Vanilla Cream Ale	/homebrew/recipe/view/1633/vanilla-cream-ale	Cream Ale	45	21.77	1.055	1.013	5.48	17.65	4.83	...	75	1.038	70.0	NaN	Specific Gravity	All Grain
2	Southern Tier Pumking clone	/homebrew/recipe/view/16367/southern-tier-pumk...	Holiday/Winter Special Spiced Beer	85	20.82	1.083	1.021	8.16	60.65	15.64	...	60	1.070	70.0	NaN	Specific Gravity	All Grain
3	Zombie Dust Clone - EXTRACT	/homebrew/recipe/view/5920/zombie-dust-clone-e...	American IPA	7	18.93	1.063	1.018	5.91	59.25	8.98	...	60	NaN	70.0	NaN	Specific Gravity	extract
4	Zombie Dust Clone - ALL GRAIN	/homebrew/recipe/view/5916/zombie-dust-clone-a...	American IPA	7	22.71	1.061	1.017	5.80	54.48	8.50	...	60	NaN	70.0	NaN	Specific Gravity	All Grain
5	Bakke Brygg Belgisk Blonde 50 l	/homebrew/recipe/view/89534/bakke-brygg-belgis...	Belgian Blond Ale	20	50.00	1.060	1.010	6.48	17.84	4.57	...	90	1.050	72.0	NaN	Specific Gravity	All Grain

5 rows × 21 columns

Imagem 1 - Fonte: Algorítimo próprio rodado no ambiente Colab.

Após a importação, realizamos uma pequena análise dos dados, obtendo o seguinte resultado:

```
[ ] print(beer_recipe.info(verbose=False))

<class 'pandas.core.frame.DataFrame'>
Int64Index: 73861 entries, 1 to 73861
Columns: 21 entries, Name to PrimingAmount
dtypes: float64(12), int64(2), object(7)
memory usage: 12.4+ MB
None
```

Imagem 2 - Fonte: Algorítimo próprio rodado no ambiente Colab.

[3] O Colab, ou "Colaboratory", é um serviço da Google que permite escrever e executar Python no navegador.

Considerando o volume das entradas, seria esperado um valor considerável de campos vazios, o que levou ao seguinte processo de descriminação de dados

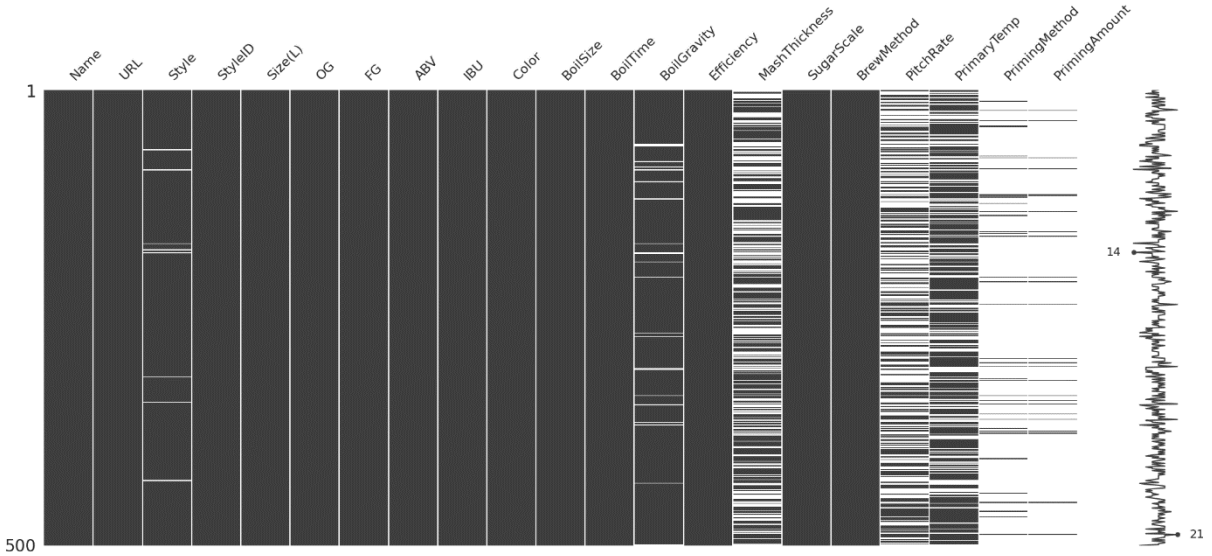

Imagem 3 - Fonte: Algorítimo próprio rodado no ambiente Colab.

Dessa forma, foi possível organizar melhor o rumo que deveríamos tomar para classificar os dados. É justamente nessa etapa que encontramos alguns problemas ignorados nas pesquisas utilizadas como referência para esse trabalho.

Apesar de ser esperado que um banco de dados desse tamanho teria diversos campos vazios, nos deparamos com uma quantidade considerável de pesquisas que ignoram a importância do Priming no processo de fabricação de cerveja. Não sabemos se isso está relacionado a possibilidade desses trabalhos encontrados terem utilizado o mesmo Dataset, ou pelo notável número de pesquisadores que admitem desconhecer tal processo.

4.2 Problema do Priming

Ao nos depararmos com tal problema, devemos fazer uma pequena crítica quanto a qualidade deste Conjunto de Dados, pois, o mesmo apresenta valores nulos ou ausentes em pontos importantes do processo de fabricação para 90% das entradas.

```
# Verificar quantidade de valores nulos na coluna 'PrimingMethod'.
null_priming = beer_recipe['PrimingMethod'].isnull()
print('PrimingMethod é nulo em {} linhas de um total de {}, representando {}%'.format(
    null_priming.sum(), len(beer_recipe), round((null_priming.sum() / len(beer_recipe)) * 100, 2)))

# Verificar quantidade de valores nulos na coluna 'PrimingAmount'.
null_priming = beer_recipe['PrimingAmount'].isnull()
print('PrimingAmount é nulo em {} linhas de um total de {}, representando {}%'.format(
    null_priming.sum(), len(beer_recipe), round((null_priming.sum() / len(beer_recipe)) * 100, 2)))
```

```
PrimingMethod é nulo em 67095 linhas de um total de 73861, representando 90.84%
PrimingAmount é nulo em 69087 linhas de um total de 73861, representando 93.54%
```

Imagem 4 - Fonte: Algorítimo próprio rodado no ambiente Colab.

A maior parte desses valores está no processo de Priming, onde, na maioria dos métodos encontrados para pesquisa e estatísticas, é possível que os dados dessa etapa sejam descartados por não serem considerados tão relevantes, ou por possuírem uma natureza mais arbitrária se comparados ao restante dos processos.

Isso acontece porque o processo de Priming é nada mais que o momento onde o Mestre Cervejeiro inicia a carbonatação, ou gaseificação da cerveja (MATOS, 2011). É preciso considerar que essa etapa representa um espaço passível de escolhas de natureza pessoal para com relação ao "gás" da cerveja, contudo, para aqueles que seguem a receita, também poderia facilmente servir como um dos itens de classificação de estilo de cerveja.

Esse é um ponto delicado para algumas das pesquisas encontradas na área, especialmente aqueles que fazem uso deste mesmo conjunto de dados, pois, embora alguns métodos considerem esses campos voláteis, essa etapa representa uma parte dinâmica do processo de produção de cerveja e está relacionado aos procedimentos pós-fermentação.

É possível que isso se dê pela abrangência de possibilidades de Priming, pois, é possível fazer uso itens como açúcar de cozinha (medido em gramas), açúcar invertido (medido em ml), capsula de CO2 (pode ser medido em unidades), etc. Cada método de Priming possuí diferentes tempos de maturação, diferentes aplicações e ainda a possibilidade de arbitrariedade em seu uso, tornando a etapa ainda mais complexa e possivelmente sendo o motivo dela não ser consenso entre fabricantes e pesquisadores (DINSLAKEN, 2015).

Independente da relevância atribuída ao processo, é certo que a carbonatação também impacta na qualidade, no tipo e na categoria da cerveja. Existem relações entre o tipo de cerveja com o método de priming, bem como o grau de CO2 a ser esperado para determinada temperatura e escada Brix. Junto a esse conceito, existem diversos guias relacionados ao tema, apresentando tabelas como seguinte:

	Volumes of CO_2										
Temp. (°F)	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**	**3.1**
33	5.0	6.0	6.9	7.9	8.8	9.8	10.7	11.7	12.6	13.6	14.5
34	5.2	6.2	7.2	8.1	9.1	10.1	11.1	12.0	13.0	14.0	15.0
35	5.6	6.6	7.6	8.6	9.7	10.7	11.7	12.7	13.7	14.8	15.8
36	6.1	7.1	8.2	9.2	10.2	11.3	12.3	13.4	14.4	15.5	16.5
37	6.6	7.6	8.7	9.8	10.8	11.9	12.9	14.0	15.1	16.1	17.2
38	7.0	8.1	9.2	10.3	11.3	12.4	13.5	14.5	15.6	16.7	17.8
39	7.6	8.7	9.8	10.8	11.9	13.0	14.1	15.2	16.3	17.4	18.5
40	8.0	9.1	10.2	11.3	12.4	13.5	14.6	15.7	16.8	17.9	19.0
41	8.3	9.4	10.6	11.7	12.8	13.9	15.1	16.2	17.3	18.4	19.5
42	8.8	9.9	11.0	12.2	13.3	14.4	15.6	16.7	17.8	19.0	20.1

Imagem 5 - Fonte: Methods of Analysis. (Milwaukee, WI: American Society of Brewing Chemists, 1949).

Desta forma, para dar continuidade a esta pesquisa, iremos considerar os campos devidamente preenchidos no Dataset mas, deixaremos registrado esse espaço para melhoria através de pesquisas futuras sobre o tema, visando compor substituições mais precisas para os valores vazios mencionados.

4.1 Categorização

Para dar início a categorização dos dados, ignoramos os valores únicos como URL, Nome, e ID. Selecionamos o campo "Style" como divisor principal, graças a sua relação com os tipos de cerveja e possíveis contribuições para organizarmos as legendas. Assim, selecionamos também os classificadores característicos do método de fabricação, sendo eles: *"OG"* (Original Gravity), *"FG"* (Final Gravity), *"ABV"* (Alcohol by Volume), *"IBU"* (International Bitterness Units), e *"Cor"*.

A seleção do campo "Style" para classificação trouxe alguns problemas relacionados a natureza dos dados disponíveis. Ao evidenciar os estilos de cerveja disponíveis no Dataset, observamos o seguinte gráfico:

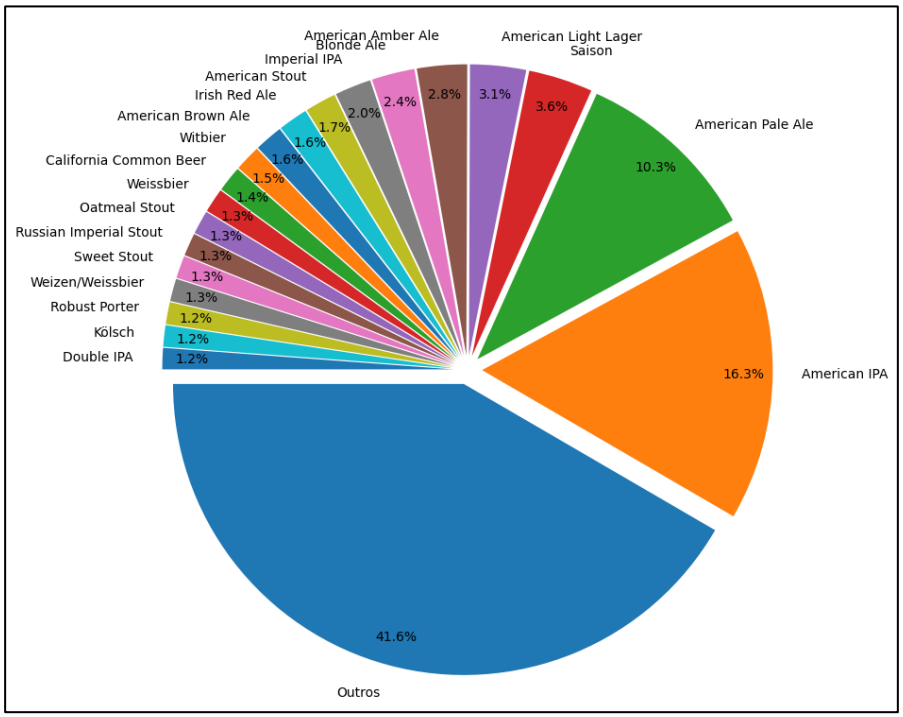

Imagem 6 - Fonte: Algorítimo utilizado para classificação

É possível notar que se considerássemos uma divisão com os 20 principais tipos de cerveja, encontramos uma grande discrepância principalmente com relação aos outros tipos restantes. Junto a isso, é possível também observar uma diferença significativa entre o tipo *"American IPA"*, *"APA"* e o restante dos estilos.

Após analisar os 176 estilos cadastrados no banco de dados do site, pudemos observar um considerável número de receitas, porém, também existem diversos campos de estilo de cerveja que são duplicatas do campo nome, ou expansões do mesmo. Isso sugere a possibilidade de uma certa confusão entre as características do método de submissão da plataforma online.

Para garantir a qualidade das simulações, fomos forçados a nos manter em 10 tipos de cerveja, pois, independentemente dos motivos para o número relativamente grande de estilos, é prudente considerar o quanto esse montante impactaria na apresentação da relação dos resultados, sendo necessário algum tipo de nivelamento e/ou padronização.

Não iremos nos aprofundar na alternativa para esse cenário devido às limitações atuais de pesquisa e a natureza simplificada deste trabalho, contudo, atestamos que estamos cientes de que para tratar tal problema, seria necessária uma completa reavaliação dos estilos de cerveja como classes, de acordo com as características de todos os campos do Dataset. Posteriormente, a aplicação de algum método de "*back-propagation"* permitiria a comparação dos estilos de cerveja a outras classes das quais eles não pertençam seria especialmente eficaz em situações onde os estilos estejam contemplando apenas uma única entrada no Dataset.

4.2 Padronização dos campos

Alguns campos representam dados em contextos de aplicação semelhantes, mas, unidades de medidas diferentes. A maior representação disso neste Dataset é a "*escala de açúcar*", sendo esta, representada em alguns casos pela unidade de medida ºP (Graus Plato) e em outros por "*SG*" (*Specific Gravity*). Para propósitos de fabricação caseira, geralmente considera-se ambas as escalas intercambiáveis enquanto tratarem de valores de até 3 casas decimais.

Para adequação de dados, utilizamos uma formula já conhecida no campo da cervejaria e providencialmente fornecida também no próprio site onde o Dataset foi gerado. Seguindo a imagem, podemos padronizar os dados numéricos do campo em questão.

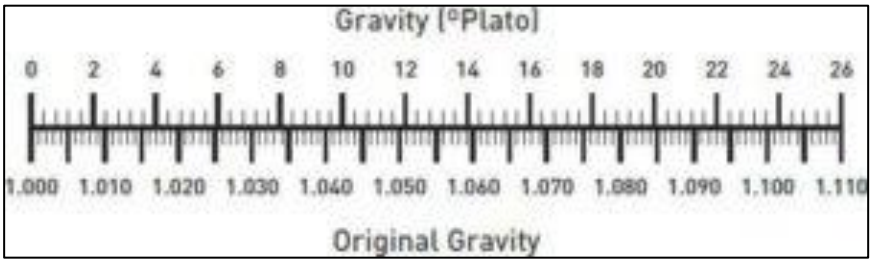

Imagem 7 - Fonte: 8degreesplato.com e brewersfriend.com

Para melhor descrever o processo, observe a seguinte formula de conversão, também disponibilizadas nas fontes mencionadas.

$$SG = 1+ (plato / (258.6 – ((plato/258.2) *227.1)))$$

$$plato = (-1 * 616.868) + (1111.14 * sg) – (630.272 * sg^2) + (135.997 * sg^3)$$

Dessa forma, para realizar as conversões em nosso algoritmo, aplicamos o seguinte trecho de código:

```
# Função para calcular SG a partir de Plato.
def get_sg_from_plato(plato):
    sg = 1 + (plato / (258.6 - ( (plato/258.2) *227.1) ) )
    return sg


# Aplicar função para criar novas colunas no DataFrame.
beer_recipe['OG_sg'] = beer_recipe.apply(lambda row: get_sg_from_plato(row['OG']) if row['SugarScale'] == 'Plato' else row['OG'], axis=1)
beer_recipe['FG_sg'] = beer_recipe.apply(lambda row: get_sg_from_plato(row['FG']) if row['SugarScale'] == 'Plato' else row['FG'], axis=1)
beer_recipe['BoilGravity_sg'] = beer_recipe.apply(lambda row: get_sg_from_plato(row['BoilGravity']) if row['SugarScale'] == 'Plato' else row['BoilGravity'], axis=1)
```

Imagem 8 - Fonte: Algorítimo próprio rodado no ambiente Colab.

Além da padronização do campo de "*escala de açúcar*", dividimos e agrupamos alguns campos de acordo com a escala de sua unidade de medida. O código a seguir contém informações sobre a disposição dos dados:

```
# Definir e Dividir campos de acordo com a escalas de suas medidas.
vlow_scale_feats = ['OG_sg', 'FG_sg', 'BoilGravity_sg', 'PitchRate']
low_scale_feats = ['ABV', 'MashThickness']
mid_scale_feats = ['Color', 'BoilTime', 'Efficiency', 'PrimaryTemp']
high_scale_feats = ['IBU', 'Size(L)', 'BoilSize']
```

Imagem 9 - Fonte: Algorítimo próprio rodado no ambiente Colab.

Considerando os dados selecionados, ponderamos a correlação entre os campos do Dataset, bem como a confiabilidade das entradas e a ausência de valores vazios (Imagem 3). Assim, como já fora mencionado anteriormente, selecionamos os campos característicos do processo de fabricação de cerveja artesanal, sendo eles: *"OG"* (Original Gravity), *"FG"* (Final Gravity), *"ABV"* (Alcohol by Volume), *"IBU"* (International Bitterness Units), e *"Cor"*.

4.3 Execução

Com todas as etapas preparadas, aplicamos um classificador rápido para testar a capacidade do algoritmo e preparar 3 métodos idealizados neste trabalho.

```
# Selecionar apenas as características relevantes.
features_list = ['StyleID', 'OG_sg', 'FG_sg', 'ABV', 'IBU', 'Color', 'SugarScale',
                 'BrewMethod', 'Size(L)', 'BoilSize', 'BoilTime', 'BoilGravity_sg',
                 'Efficiency', 'MashThickness', 'PitchRate', 'PrimaryTemp']
clf_data = beer_recipe.loc[:, features_list]

# Label encoding para características categóricas.
cat_feats_to_use = list(clf_data.select_dtypes(include=object).columns)
for feat in cat_feats_to_use:
    encoder = LabelEncoder()
    clf_data[feat] = encoder.fit_transform(clf_data[feat])

# Preencher valores nulos para características numéricas.
num_feats_to_use = list(clf_data.select_dtypes(exclude=object).columns)
for feat in num_feats_to_use:
    imputer = SimpleImputer(strategy='median')
    clf_data[feat] = imputer.fit_transform(clf_data[feat].values.reshape(-1, 1))
```

Imagem 10 - Fonte: Algorítimo próprio rodado no ambiente Colab.

Além de selecionar os campos de maior relevância, foram também atribuídos valores nulos para características numéricas. Isso foi necessário para trabalhar corretamente o método "*XGBClassifier"*. Feito isso basicamente a próxima tarefa é apenas a execução dos métodos classificadores "*Random Forest Classifier"*, "*Logistic Regression"* e "*XGBoost Classifier"*, sendo estes separados por caractere de comentário, impedindo que todos executem ao mesmo tempo e causem "*overflow*" no ambiente.

```
# Escolher classificador (descomentar um deles).

# clf = RandomForestClassifier()
# clf = LogisticRegression()
clf = xgb.XGBClassifier()
```

Imagem 11 - Fonte: Algorítimo próprio rodado no ambiente Colab.

4.4 Resultados

Com a utilização de cada um dos métodos, foi possível realizar previsões quanto as classificações dos estilos das receitas de cerveja e demonstrar os campos mais importantes utilizados nessas previsões. Infelizmente nenhum dos métodos conseguiu atingir um alto grau de precisão, chegando à eficácia máxima de 37,6% para o "*XGB.Classifier*", seguido por 32,1%

para "*Random Forest Classifier"* e por fim "*Logistic Regression"* com 22,5%, apresentando o pior resultado dos três métodos utilizados.

Ambos os 3 métodos devolveram resultados semelhantes com relação aos campos mais importantes do Dataset, permitindo realizar uma pequena comparação com a realidade comum do âmbito de fabricação de cerveja. Esses resultados podem ser conferidos no gráfico a seguir:

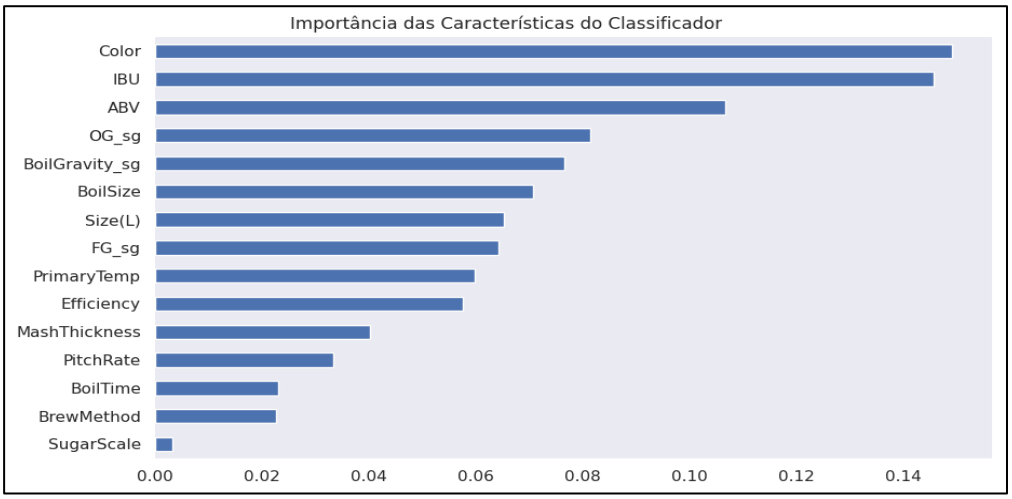

Imagem 12 - Fonte: Algorítimo próprio rodado no ambiente Colab.

De acordo com os resultados mencionados, podemos observar como principais características de classificação a "*cor*", o "*IBU*" (International Bitterness Unitis) e o "*ABV"* (Alcohool By Volume). Esses dados refletem a realidade do consumidor com relação a escolha de sua cerveja.

A cor, o teor alcoólico e o amargor são considerados os fatores mais importantes da bebida, sendo a cor um indicador da variedade a ser consumida, enquanto, o teor alcoólico também está diretamente associado a isso.

MACRO DIVISÃO	SRM	TONALIDADE	EBC	CLASSIF.**
Palha	2 – 3		3,94 – 5,91	
Amarelo	3 – 4		5,91 – 7,88	Cerveja Clara até 20 EBC
Ouro	4 – 5		7,88 – 9,85	
Âmbar	6 – 9		11,82 – 17,73	
Profundo âmbar / cobre luz	10 – 14		19,70 – 27,58	
Cobre	14 – 17		27,58 – 33,49	
Profundo cobre/castanho claro	17 – 18		33,49 – 35,46	
Castanho	19 – 22		37,43 – 43,34	Cerveja Escuro ≥ 20 EBC
Castanho Escuro	22 – 30		43,34 – 59,10	
Castanho muito escuro	30 – 35		59,10 – 68,95	
Preto	35 +		68,95 – 78,80	
Preto opaco	40+		>78,80	

Imagem 13 - Fonte: Tabela de cores da cerveja artesanal – Revista Malagueta.

Cada variedade de cerveja é reconhecida por seus consumidores através de características marcantes que são refletidas nesses detalhes e compreendidas pelos diversos fatores sensoriais daqueles que as apreciam.

4.3 Discussão

Neste estudo, exploramos um conjunto de dados que contém 73861 receitas de cerveja. Embora essa quantidade seja considerada relativamente abrangente, é importante lembrar que, em termos de dados, essa quantidade é consideravelmente limitada. No entanto, essa limitação não deve ser vista como uma desvantagem para este trabalho em questão. A quantidade restrita de dados contribui para a eficiência computacional dos algoritmos, permitindo que eles sejam executados em ambientes virtuais de processamento, como a plataforma Colab da Google.

Ao importar o arquivo contendo o Dataset, os dados foram reestruturados e transferidos para um sistema de hospedagem personalizado. A análise dos dados revelou um problema crítico relacionado à presença de valores nulos ou ausentes em pontos cruciais do processo de fabricação, afetando cerca de 90% das entradas. Para lidar com essa lacuna de dados, preenchemos os valores faltantes com as entradas mais frequentes, possibilitando aumentar a integridade dos conjuntos de dados.

Um desafio adicional emergiu durante a análise dos dados, especificamente relacionado ao processo de Priming. Ficou evidente que os dados relacionados a esse processo eram frequentemente inexistentes. O processo de Priming é crucial para a carbonatação da cerveja, no entanto, muitas pesquisas não dão a devida importância a esse estágio ou, em alguns casos, até desconhecem sua existência. Esse problema coloca em questionamento a qualidade dos dados presentes no conjunto.

Para permitir a categorização dos dados, selecionamos o campo "Style" como principal divisor, juntamente com outros classificadores representativos das características do processo de fabricação. Além disso, foi necessário padronizar campos que apresentavam dados em diferentes unidades de medida, como a escala de açúcar por exemplo. Embora a padronização tenha sido eficaz, vale destacar que outras abordagens poderiam ter sido exploradas, principalmente em outros campos como a conversão de todas as unidades para um padrão mais eficaz, a fim de simplificar a análise.

Com os dados preparados, procedemos à aplicação de três métodos de classificação: “*Random Forest Classifier”*, “*Logistic Regression”* e “*XGBoost Classifier”*. Os métodos foram treinados no conjunto de treinamento e avaliados no conjunto de testes. No entanto, os

resultados não alcançaram as expectativas estabelecidas. A precisão máxima atingida foi de apenas 37,6% pelo XGBoost Classifier.

Assim, estudo evidenciou os desafios encontrados ao trabalhar com o conjunto de dados e a aplicação dos métodos de classificação. Apesar dos resultados não serem os desejados, eles fornecem valiosos insights sobre a qualidade dos dados e a complexidade da classificação das receitas de cerveja. Além disso, tal fato ressalta a importância de dados de alta qualidade e a necessidade de considerar fatores específicos do processo de fabricação e sua influência no mesmo. Concomitantemente, uma abordagem mais robusta de limpeza de dados poderia ter sido explorada, incluindo a substituição de valores ausentes por estimativas mais precisas ou a realização de análises de sensibilidade para avaliar o impacto de diferentes abordagens de preenchimento.

Ainda assim, é importante destacar que, apesar dos desafios, este trabalho lança luz sobre a complexidade escalável em analisar conjuntos de dados tão heterogêneos, revelando oportunidades para futuras pesquisas que visem a melhora da qualidade dos métodos de classificação utilizados.

5 – CONSIDERAÇÕES.

Este trabalho descreveu os paços de uma análise detalhada de um extenso conjunto de dados contendo 73.861 receitas de cerveja artesanal. Nosso objetivo principal foi aplicar métodos de aprendizado de máquina para classificar essas receitas de acordo com seus estilos permitindo, contribuir para previsões mais precisas durante o processo de fabricação. Ao longo desse estudo, enfrentamos desafios significativos e obtivemos informações cruciais para a expansão do tema e ciência do perfil pedagógico esperado na aplicação dos métodos utilizados.

Abordamos também a importância econômica e cultural da cerveja, especialmente da cerveja artesanal, que tem ganhado um espaço significativo no mercado brasileiro e mundial. Ainda assim, observamos que as pesquisas nesse domínio muitas vezes carecem de informações detalhadas sobre o processo de fabricação, levando a lacunas de dados e outros problemas.

Enfrentamos o desafio de lidar com valores ausentes, principalmente no que diz respeito ao processo de Priming, uma etapa crucial para a carbonatação da cerveja. Essa lacuna de dados nos fez questionar a qualidade do conjunto de dados e ressaltou a necessidade de pesquisas futuras para melhorar a completude das informações relacionadas a esse processo.

A categorização das receitas foi uma etapa crítica, pois, optamos por utilizar o campo "Style" como o principal divisor, contudo, identificamos problemas de duplicação e

inconsistências nesse campo, o que impactou nossa análise. Uma abordagem mais aprofundada de classificação dos estilos de cerveja será considerada em pesquisas futuras.

Mostrou-se necessária a padronização dos dados, especialmente as unidades de medida, para garantir que os algoritmos de aprendizado de máquina pudessem funcionar da forma mais eficaz possível. Dessa forma, reconhecemos que outras abordagens de normalização ou conversão de unidades poderiam ser mais amplamente exploradas, ainda que deliberadamente tenhamos optado por não o fazer devido as características limitadas da natureza contextual desse trabalho.

A aplicação dos três modelos de classificação "*Random Forest Classifier"*, "*Logistic Regression"* e "*XGBoost Classifier"*, infelizmente, não atendeu às nossas expectativas de precisão. Os melhores resultados foram alcançados pelo "*XGBoost Classifier"*, com uma precisão máxima de 37,6%, o que se alinha a esperada a complexidade da tarefa de classificação das receitas de cerveja com base em seus atributos.

Em suma, este estudo ofereceu uma visão detalhada dos desafios enfrentados ao lidar com conjuntos de dados no domínio da cerveja artesanal e ao aplicar métodos de aprendizado de máquina para classificação. Embora os resultados não tenham sido ideais, eles fornecem uma base sólida para pesquisas futuras e destacam a importância de dados de alta qualidade. Além disso, o proposito pedagógico relacionado aplicação de métodos de classificação, aprendizado de máquina e inteligência artificial, mostrou-se de fundamental importância para o mundo atual e a ciência contemporânea. Dessa forma, contribui para a compreensão aprofundada do processo de fabricação, do tratamento de conjuntos de dados mais densos, como os relacionados à produção de cerveja, além da aplicação de métodos de classificação e aprendizado de máquina para o setor cervejeiro.

A partir deste ponto, diversas direções podem ser seguidas para aprimorar os conhecimentos sobre o tema e melhorar a qualidade das previsões. Ainda que tenhamos apenas "arranhado" a superfície deste vasto tema, este estudo será crucial para direcionar nossos próximos passos na área, já idealizando o próximo projeto norteado pelo preenchimento mais preciso dos valores ausentes e a exploração de novas características junto a implementação de abordagens mais ainda mais avançadas de aprendizado de máquina.

Com um conjunto de dados aprimorado e metodologias mais sofisticadas, é possível realizar análises muito mais elaboradas, contribuindo ainda mais para o mundo e para a ciência, inclusive a ciência da produção de cerveja artesanal.

REFERÊNCIA BIBLIOGRÁFICA

CERVBRASIL (2018). **DADOS DO SETOR CERVEJEIRO NACIONAL**. Disponível em: <www.cervbrasil.org.br/novo_site/dados---do---setor>. Acessado em 30 de agosto de 2023.

CHEN, T., & GUESTRIN, C. (2016). **XGBoost: A scalable tree boosting system**. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

DIAS, M. O. (2018). **Indústria cervejeira da Heineken no Brasil. In: International Journal of Management, Technology And Engineering** (IJAMTES) ISSN: 2249-7455. Volume 8 Edição 9, novembro/2018, Página nº: 1304-1310. DOI:16.10089/IJMTE2156

DINSLAKEN, D. **Manual do cervejeiro caseiro: Um guia completo para iniciantes.** CONCERVEJA - Congresso nacional de cerveja. 2ª Edição. 2015.

MATOS, R. G. M. **Produção de Cervejas Artesanais, Avaliação de Aceitação e Preferência, e Panorama do Mercado.** 2011. 78p. TCC (Graduação em Agronomia), UFSC – Universidade Federal de Santa Catarina. Florianópolis - SC. Disponível em: < https://repositorio.ufsc.br/xmlui/handle/123456789/25472>, Acessado em: 24/07/2023.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2013). **An Introduction to Statistical Learning**. Springer.

LIAW, A., & WIENER, M. (2002). **Classification and Regression by randomForest**. R News, 2(3), 18-22.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2009). **The elements of statistical learning: data mining, inference, and prediction**. Springer Science & Business Media.

HOSMER JR, D. W., & LEMESHOW, S. (2004). **Applied *Logistic Regression***. John Wiley & Sons.

TRAPPEY, A. J., TRAPPEY, C. V., TRAPPEY, A. J. C., KU, Y. C., & GOVINDAN, K. (2016). **The fuzzy AHP approach for developing ingredient intelligence in cooking using ontologies**. Journal of Food Engineering, 175, 99-110.

TUKEY, J. W. (1977). **Exploratory data analysis.** Addison-Wesley.