# Introdução

* 1. *Justificativa*

A cafeicultura é representativa na produção agrícola de países aptos ao cultivo deste grão na América do Sul. Nesses mercados, a produção pode ser distinguida entre os lotes de melhor qualidade, destinados à exportação aos grandes centros como Europa e Estados Unidos, e os que não atingem a qualidade necessária para este fim, usados para suprir a demanda local. O consumo deste produto é crescente a nível mundial, especialmente pela entrada de mercados significativos antes dominados pelo consumo de chá, a exemplo da China, que passa por um processo de aproximação cultural com o ocidente e, assim, absorve lentamente costumes antes inexistentes naquela região (MONTE, D., LOPES, D., CONTINI, E 2017). Brasil e Colômbia disputam a liderança da produção e qualidade com países da África Oriental e do Sudoeste Asiático.

Há um particular interesse na melhoria da produção na América do Sul, a fim de conseguir melhores preços no mercado internacional. Usualmente, a produção é escoada por intermediários que fazem a seleção de lotes aptos à exportação junto aos produtores, em boa parte operando em escala familiar e artesanal. Na Colômbia, estima-se que sua produção total seja composta por cerca de 560.000 pequenos produtores, segundo Salamanca, 2015. Existem empresas especializadas em avaliar os lotes, levantar quantidade de produto, negociar com compradores do exterior e enviar as remessas, estando responsáveis pela qualidade do café que chega aos centros consumidores.

A qualidade do café é medida por um método desenvolvido pela Speciality Coffe Association (SCA), denominado Protocolo SCA, podendo ser aplicado por qualquer parte que obtenha o certificado fornecido pela associação. Consiste em uma gradação de zero a cem pontos denominada por dez atributos classificados de seis a dez, classificado conforme tabela abaixo:

Tabela

Descrição gerada automaticamente

Tabela 1: Classificação de qualidade de cafés segundo Protocolo SCA

Trata-se do padrão mundial para avaliação de café, apesar de existirem outros protocolos com a mesma finalidade. Os atributos que compõe o protocolo são:

1. Fragrância e aroma, onde o primeiro refere-se ao cheiro do pó seco recém-moído e o segundo à bebida já umedecida.
2. Sabor, referente ao gosto da bebida e suas notas
3. Retrogosto, a sensação que fica após o consumo do líquido
4. Acidez, referente ao tipo e intensidade da acidez
5. Corpo, a quantidade de óleos essenciais na bebida
6. Uniformidade, medida ao analisar cinco xícaras diferentes do mesmo café
7. Ausência de defeitos, que avalia os sabores estranhos da bebida
8. Doçura mínima, variável de acordo com a torra do café
9. Balanço, a junção de sabor, acidez e corpo
10. Conceito final, que representa a impressão do provador sobre o café

Caso algum dos atributos receba nota inferior a seis, o café já não é considerado especial. Segundo entrevista com negociantes colombianos, em geral cafés com nota superior a 85 são destinados à exportação sob valor superior.

A nota SCA é a avaliação final do produto derivado de diversos processos agronômicos e de pós-colheita. O meio que o produtor possui de elevar a nota dos seus cafés é melhorar esses processos, de forma que reflita nas características sensoriais dos lotes.

Os comerciantes de café compram os lotes sem ter a nota SCA dos mesmos, uma vez que a nota é atribuída na etapa de exportação. No entanto, é possível coletar atributos durante as etapas pós-colheita do café que estão correlacionadas com a nota final que o lote obterá.

Dessa forma, levanta-se a hipótese de gerar um modelo capaz de predizer a nota SCA de um lote a partir de determinadas características coletadas. Isso seria útil para a tomada de decisão dos intermediários na hora de efetuar a compra de um lote. Ao mesmo tempo, seria possível premiar produtores capazes de proporcionar cafés especiais com melhores preços de compra. Dado o caráter subjetivo e empírico da avaliação (THOMAS, E., PUGET, S., VALENTIN, D., SONGER, P. 2017), entende-se que pode haver erros no processo. Dessa forma, a utilização de ferramentas estatísticas é uma forma de aumentar a eficiência e confiabilidade das avaliações

* 1. *Problema de pesquisa*

A Colômbia é um país relevante na cafeicultura mundial desde o século XIX, tendo tradicionalmente produzido cafés especiais. O regime de chuvas e temperatura é favorável ao cultivo da espécie Arábica *(Coffea Arabica*), mais propensa a originar bebidas de alto padrão (ARCILA, 2007). Atualmente, a Colômbia é o terceiro maior produtor mundial deste grão (THOMAS et al., 2017), representando de 5% a 7% das exportações totais do país (DE OLIVEIRA, E.M., LEME, D.S., BARBOSA, B.H.G., RODARTE, M.P. 2105).

A demanda é oriunda de comerciantes de café colombiano interessados em obter tal modelo. Para tanto, foi fornecido uma base com observações de lotes de cafés com suas respectivas notas SCA.

Esses lotes foram monitorados através de etapas de processamento do café colhido. Inicialmente a recepção do lote no armazém, passando pela despolpa, fermentação, lavagem, secagem até o armazenamento do grão antes da torrefação. Durante esse processo mediu-se pH, temperatura, teor de açúcar, umidade e o tempo que o lote esteve em cada etapa. Os lotes diferem entre si por sua origem e variedade de café, no entanto o presente estudo foca nas variáveis quantitativas coletadas durante o processamento.

Sendo assim, levanta-se questões a respeito da qualidade da base de dados. Inicialmente, os campos foram avaliados e houve uma tentativa de melhorar a coleta dos dados afim de gerar uma base mais robusta. A partir dessa base, esta pesquisa inicialmente visará obter a melhor maneira de tratar os dados para posteriormente concluir qual método mais eficaz para o treinamento de um modelo preditivo capaz de estimar a nota SCA de um lote de café usando como insumo atributos físicos e químicos da produção

* 1. *Objetivos*

Comparar formas de pré-processamento da base de dados bem como métodos de treinamento de modelo preditivo. Serão utilizadas métricas tradicionais para avaliar os modelos, permitindo assim eleger um fluxo de treinamento ideal para a base de dados em questão. Também procura-se entender quais são as variáveis mais relevantes para o treinamento do modelo e, portanto, as mais correlacionadas

# Revisão Bibliográfica

O estudo de Ribeiro, Augusto, Salva e Ferreira (2012) é destaque no âmbito da modelagem preditiva de qualidade de cafés na América do Sul. O estudo destaca o fato de protocolos similares ao SCA poderem ser considerados subjetivos e demorado, o que possivelmente seria um empecilho para um modelo de aprendizagem de máquina, ao mesmo tempo que é motivo de um modelo como esse ser necessário.

O artigo difere do presente estudo por tratar-se de cafés já torrados. Um grupo de componentes análogos aos encontrados no protocolo SCA foram utilizados como variável alvo, sendo estes os seguintes: acidez, amargor, sabor, limpeza, corpo e qualidade geral. Para tanto, experts efetuaram teste de análise sensorial a atribuíram notas à cada componente para as amostras. Este conjunto foi posteriormente tratado para que se chegasse em um valor para cada amostra e atributo.

Atributos quimiométricos extraídos por análise cromatográfica foram utilizados como variáveis independentes. O método *Partial Least Squares* (PLS) foi utilizado para gerar a regressão, calibrados por 43 amostras. Cada um dos atributos de bebida analisados teve um modelo particular, contando com um conjunto diferente de variáveis dentre as 20.640 levantadas. Nenhuma amostra foi descartada classificada como *outlier*.

O erro observado no conjunto de 10 observações teste foi similar aos erros atribuídos às avaliações de *experts* no assunto. Concluiu-se que é possível predizer as notas dos atributos componentes do protocolo SCA utilizando dados provenientes da cromatografia de uma amostra através de um modelo de regressão. Esse modelo é capaz de atribuir classificação semelhante à dos profissionais capacitados a atribuí-las. Finalmente, se comparado os métodos de extração de variáveis via coromotgrafia com o método de espectrometria usando ondas infravermelho próximo, o conjunto de dados obtidos via cromatografia foi capaz de gerar um modelo mais eficaz.

O estudo de Tolessa, Rademaker, De Baets e Boeckx datado de 2015 é um exemplo de modelagem baseada em um conjunto de dados obtidos por meio de espectrometria de ondas de infravermelho próximo em cafés crus produzidos na Etiópia. É destacado o fato de tratar-se de um método rápido e barato, dado que não demanda nenhum processo químico e pouco preparo sobre as amostras, bem como pouca manutenção de aparelhagem, sendo assim uma técnica de fácil replicação na indústria.

Recorrentemente, a espectrometria é utilizada para distinguir cafés do tipo Arábica de cafés do tipo Robusta, cafés maduros de verdes e grãos defeituosos de não defeituosos. No entanto, o artigo destaca que a maioria dos estudos que provam a correlação de dados de espectrometria com diferentes classificações de lotes de café se deram com cafés torrados, não havendo iniciativas desse tipo com café cru.

Os dados obtidos foram tratados a fim de minimizar os ruídos. Para gerar o modelo, obteve-se os dados a partir de 86 amostras de café maduro colhido em diferentes localidades do país, sendo posteriormente secados e lavados para serem submetidos à espectrometria. Também se utilizou o método PLS para obtenção da regressão do conjunto de dados com os atributos de qualidade de bebida, sendo as observações distinguidas em 66 para treino e 20 para validação do modelo.

Os modelos foram avaliados usando a raiz quadrática média do erro da validação cruzada, a raiz quadrática média do erro da validação externa (predição), os resíduos do desvio de predição e o coeficiente de correlação (r).

O estudo obteve resultados promissores no que concerne a qualidade geral do café e os demais atributos de bebida avaliados, chegando em um valor r = 90 para correlação entre os valores obtidos pelo modelo e os valores segundo a avaliação de especialistas no caso de lotes de alta qualidade. Contudo, o modelo teve problemas para classificar cafés de qualidade inferior, obtendo valor r = 59 entre os valores de qualidade geral obtidos pelo modelo e os valores segundo avaliação de especialistas. Isso explica-se pela dificuldade de captar a presença de defeitos nos grãos via espectrometria, defeitos estes mais comuns em cafés de menor qualidade.

O artigo de J.S. Ribeiro et al. (2010) vem em consonância com Tolessa et al., contribuindo para a conclusão de que a espectrometria é um método válido para obtenção de dados capazes de gerar modelos de predição de qualidade de café. Este estudo contou com 51 amostras, colhidas e processadas no Brasil. Este difere do presente estudo por basear-se em amostras já torradas. Do mesmo modo que as pesquisa previamente descritas, também foi utilizado o método PLS para modelagem. Os mais de 700 espectros de reflectância obtidos foram separados para a criação dos modelos para cada um dos seis atributos de bebida estudados. As observações foram separadas em 41 casos como conjunto de treino e 10 casos como conjunto de teste. Afinal, a pesquisa obteve resultados inferiores à Tolessa et al., no entanto isso pode ser advindo do fato de ter tido menos lotes de café disponíveis para o treinamento do modelo.

Finalmente, destaca-se o artigo de Suarez-Peña, Lobaton-García, Rodríguez-Molano, Rodriguez-Vazquez, (2020), regionalmente focado na Colômbia, assim como esse estudo. O objetivo foi obter uma modelagem estatística de qualidade de café baseada em características físico-químicas que fossem possíveis de serem coletadas por uma ampla gama de profissionais em diferentes regiões em cafés crus, tendo assim um modelo generalista e replicável. Para tanto, foram levantadas características inerentes às amostras, como peso, umidade, cor e presença de impurezas. A classificação objetivada seria binária, sendo as amostras classificadas apenas como boas ou ruins.

O conjunto de dados foi normalizado, de modo que posteriormente seria feita a comparação da acurácia dos modelos gerados à partir do conjunto de dados original e do naormalizado. Foi feito uma análise de correlação entre as variáveis independentes e a classificação das amostras via análise sensorial. O estudo comparou os resultados das modelagens obtidos via redes neurais e *Support Vector Machine* (SVM).

A rede neural obtida teve como insumo 36 nós e desenvolve-se a partir de 4 camadas, atingindo acurácias de 83%. Notou-se um melhor desempenho do modelo para amostras classificadas como ruins, tendo 100% de acerto, em comparação com 67% das amostras consideradas boas. Após submetido a validação cruzada, o modelo atingiu acurácia geral de 81%, ainda com melhor desempenho em amostras ruins. Notou-se que as amostras classificadas erroneamente tiveram nota próxima ao limiar das classificações, de modo que com a inclusão de novas classes poderia ter sido obtido um modelo mais assertivo.

O método SVM foi aplicado com diferentes calibrações de parâmetros, utilizando-se posteriormente da mais eficaz. Após a validação cruzada, obteve-se acurácia geral de 64%. Não houve diferença nos resultados obtidos pelo conjunto de dados original e normalizado, bem como após a aplicação de *Principal Components Analysis* (PCA), a fim de reduzir a dimensionalidade do conjunto. O método SVM demandou menor tempo de processamento para ser gerado se comparado com as redes neurais.

Os autores sugerem que uma possível melhoria para futuros estudos seria, além de testar novos algoritmos de treinamento de modelo, combinar medições vindas de atributos físico-químicos com fotografias padronizadas das amostras como insumo de algoritmo de redes neurais convolucionais.

# Material e Métodos

Para este estudo, foi disponibilizada uma base com 1002 observações de lotes de café comercializados na Colômbia. Inicialmente, a base possuía 71 campos, contendo informações coletadas de cada lote em sua fase de recepção, fermentação, despolpa, fermentação, lavagem e secagem. As observações datam de maio de 2019 a maio de 2022. Os campos consistem em dado de pH, concentração de açúcar, temperatura da amostra, peso, umidade e minutos em que ficou em cada etapa do processamento. Posteriormente, gerou-se novas informações com as diferenças de pH, concentração de açúcar e temperatura entre as diferentes etapas. A tabela a seguir, ilustra a confiabilidade de cada campo, medida pela porcentagem de valores faltantes e de valores discrepantes, utilizando como base valores entre o primeiro e o terceiro quartis do campo.

Gráfico, Histograma

Descrição gerada automaticamente

Figura 1: confiabilidade dos campos componentes da base fornecida

Foi feito uma tabela de correlação entre todos os campos e a variável dependente, onde verificou-se que nenhum deles tinha correlação evidente, indicando que uma abordagem utilizando correlação linear não seria efetiva. Da mesma forma, não se verificou sazonalidade na qualidade das amostras, utilizando como base a data de recepção do lote

Gráfico

Descrição gerada automaticamente

Figura 2: Correlação entre variáveis preditoras e variável dependente

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figura 3: Média de qualidade das amostras por mês de recebimento

Sendo assim, decidiu-se por aplicar técnicas de aprendizado de máquina afim de obter o modelo de predição da qualidade dos lotes. As observações foram distinguidas em 5 classes: amostras superiores à 90 de SCA foram classificadas como “Excepcionais”, as superiores à 80 como “Excelentes”, as superiores à 70 como “Muito boas” e por fim as superiores à 60 como “Boas”. As que não atingiram essa marca foram classificadas como “Insuficientes”. A base foi separada em treino e teste na proporção de 75% e 25% respectivamente e preencheu-se valores faltantes com a mediana da coluna correspondente. Para efeito de comparação, utilizou-se 5 algoritmos diferentes para a classificação: árvore decisória, random *forest, K-nearest neighbours, support vector machine* e *XGBoost*.

# Resultados e Discussão

# Referências

MONTE, D., LOPES, D., CONTINI, E.. China: Nova potência também no agronegócio. **Revista de Política Agrícola**, 26, Dez. 2017. Disponível em: <<https://seer.sede.embrapa.br/index.php/RPA/article/view/1301>>. Acesso em: 11 Dez. 2022.

ARCILA, Jaime et al.: Sistemas de producción de café en Colombia. 2007.

RIBEIRO, J. S. et al.: Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics. Talanta, v. 101, p. 253-260. 2012.

THOMAS, E., PUGET, S., VALENTIN, D., SONGER, P.: Sensory Evaluation-Profiling and Preferences. Elsevier Inc., London. 2017.

SALAMANCA, C.: Métodos Estadísticos Para Evaluar La Calidad Del Café. Universitat de Girona. 2015.

DE OLIVEIRA, E.M., LEME, D.S., BARBOSA, B.H.G., RODARTE, M.P.: A computer vision system for coffee beans classification based on computational intelligence techniques. 2015.

SUAREZ-PEÑA, J.A., LOBATON-GARCÍA, H.F., RODRÍGUEZ-MOLANO, J.I., RODRIGUEZ-VAZQUEZ, W.C.: Machine Learning for Cup Coffee Quality Prediction from Green and Roasted Coffee Beans Features. 2020.

TOLESSA, K., RADEMAKER, M., DE BAETS, B., BOCEKX, P.: Prediction of specialty coffee cup quality based on near infrared spectra of green coffee beans,Talanta, Volume 150, pag. 367-374. 2016.