Predição de qualidade do vinho

Diego de Almeida Miranda - 133603

# Introdução

O mercado de vinho no mundo movimenta milhões de dólares e atinge diversos tipos de consumidores, gerando uma disponibilidade de produtos mais acessíveis e populares, e também produtos mais luxuosos. Os principais exportadores de vinho no mundo são países Europeus que totalizam anualmente 10 milhões de toneladas de vinho, segundo a Food and Agriculture Organization of the Nations (FAO) em 2013. Ainda segundo os dados da FAO, os principais países exportadores no mundo são sequencialmente a Itália, Espanha, França, Chile e Austrália. O Brasil, por sua vez, importa cerca de 72 mil toneladas de vinho anualmente, em contraste às 20 mil toneladas da bebida que é exportada. Com esse valor, o Brasil está entre os 30 países que mais exportam a bebida para outros países.

Diversas tecnologias têm sido desenvolvidas e aplicadas, não só para a produção de vinhos, mas também para sua comercialização. A certificação dos vinhos é um dos processos de suma importância neste comércio. É a partir da certificação que podemos ter o controle de qualidade e segurança do produto. A certificação previne possíveis alterações ilegais na produção do vinho, como também traz prestígio para as marcas.

A certificação de um vinho é comumente feita após um diagnóstico das propriedades físico químicas e sensoriais do produto. As propriedades físico químicas normalmente analisadas são o pH, o nível de açúcar no produto final, a densidade do vinho, etc. Por outro lado, as qualidades sensoriais da bebida só podem ser avaliadas através da degustação de pessoas qualificadas, experts, na área. Ainda é objeto de estudo científico de que maneira as propriedades físico-químicas interferem nas análises sensoriais do vinho. Ainda assim, sabemos que, por exemplo, o ácido cítrico pode adicionar frescor e sabor aos vinhos

Com o desenvolvimento tecnológico foi tornando-se cada vez mais possível a coleta, o armazenamento e o processamento dos dados físico químicos e sensoriais obtidos no processo de certificação de uma bebida. Os dados aqui apresentados são referentes aos vinhos da empresa Portuguesa "Vinho Verde". Os dados que indicariam quais tipos de uva foram usadas ou à qual valor os vinhos são vendidos não foram disponibilizados por questões éticas e de privacidade da empresa. Ainda assim, a partir dos dados presentes neste conjunto podemos realizar análises suficientemente eficientes para o fim de estudo e compreensão do processo de modelagem de dados.

Através de métodos estatísticos, de aprendizado de máquina e análise exploratória, podemos realizar a então chamada mineração de dados, a qual busca compreender e descrever os padrões ocultos "a olho nu". Isto é, através de ferramentas computacionais podemos realizar análises em coleções de dados exorbitantemente grandes para uma pessoa, ou até mesmo um grupo de pessoas, analisar(em). O grande volume de dados é um aspecto crescente em todas as áreas do conhecimento, e utilizar ferramentas que consigam processar todo esse volume de uma maneira eficiente e coerente é essencial para o desenvolvimento científico e capital. Empresas das mais variadas conseguem gerar Terabytes de dados sobre seu serviço sem muito esforço, tendo isso em vista, aplicar métodos computacionais para encontrar informações "escondidas" nos dados é algo de suma importância.

No caso dos vinhos, podemos utilizar a computação para buscar compreender quais os atributos e propriedades físico-químicas estão relacionadas a um vinho de boa qualidade. Por exemplo, podemos buscar saber através de numerosas amostras de vinho qual a densidade ideal de uma bebida avaliada com notas a partir de 7. Podemos especular algumas perguntas a serem respondidas através dessa modelagem de dados. Por exemplo, neste conjunto de dados, quais os atributos físico-químicos mais relevantes na classificação de um bom vinho? Dentre estes atributos importantes, quais os valores ideias para esses atributos? Como esses atributos estão relacionados entre si? Qual o melhor modelo para predição de um bom vinho?

Serão utilizados modelos de aprendizado de máquina supervisionado, isto é, modelos que precisam ser ajustados através de um conjunto de amostras rotulados para que posteriormente possa realizar a predição de um determinado conjunto de amostras. No geral esse processo de predição é feito entre duas classes distintas, mas frequentemente encontramos problemas de predição multi-classes, o que torna o problema mais complicado de ser resolvido. Os modelos de aprendizado de máquina selecionados para resolver o problema de classificação foram os baseados em Support Vector Machines, K-nearest neighbours e Random Forests.

É natural que, ao dividirmos o conjunto de dados em dados para teste e em dados para treino, criemos um modelo enviesado e seus resultados não sejam de alta confiabilidade. Isto claro, a depender da qualidade e quantidade de dados inseridos no modelo. Ainda assim, formas diferentes de particionar o conjunto de dados podem gerar resultados e valores de métricas diferentes. Com o intuito de obtermos resultados mais confiáveis nas métricas de avaliação, utilizaremos do método de validação cruzada. A validação cruzada consiste em participar do conjunto de dados em k subconjuntos, sendo k-1 utilizados para treino e o subconjunto restante utilizado para teste deste modelo. A partir do resultado médio dessas métricas podemos ter uma noção mais confiável do real desempenho dos modelos neste conjunto de dados.

Muito foi dito sobre as métricas de avaliação a serem utilizadas, mas não foram ditas quais dessas serão usadas. Os modelos aqui apresentados serão avaliados utilizando as métricas de acurácia, área sob a ROC - conhecido como AUC ROC - e pontuação F1. A acurácia é dada pela divisão entre a soma dos resultados verdadeiros (positivo e negativo) pela quantidade de amostras que temos. Sendo \*VP\* amostras classificadas corretamente como positivas, e \*VN\* amostras classificadas corretamente como negativas, então a acurácia \*acc\* pode ser dada por $acc = \frac{VP + VN}{|amostras|}$. A métrica de pontuação F1 é a média harmônica entre precisão e revocação, e é dada pela seguinte equação $f1 = 2\frac{Precisao\times Revocaçao}{Precisao + Revocaçao}$. Por fim, a curva ROC é realizada através das taxas de verdadeiro positivo e verdadeiro negativo. No eixo X da curva ROC temos o \*False positive Rate\*, onde $FPR = \frac{FP}{VP + FP}$. De forma análoga, temos no eixo Y temos o \*True Positive Rate\*, dado por $FPR = \frac{VP}{VP + FN}$. A métrica de AUC nada mais é então que o valor da área sob a curva ROC.

Todos os modelos e ferramentas necessárias para sua avaliação e validação foram obtidas através das bibliotecas em Pyhon3 Numpy, Pandas, Matplotlib e, principalmente, sci-kit learning.

# Análises

As análises podem ser verificadas no pdf associado a este trabalho.

# Conclusão

Agora que já foram feitas as implementações gerais da modelagem, é necessário que façamos uma análise sobre os resultados que foram obtidos. É importante que tenhamos respondido as questões levantadas no início deste trabalho. Repito aqui as perguntas que foram levantadas: quais os atributos físico-químicos mais relevantes na classificação de um bom vinho? Dentre estes atributos importantes, quais os valores ideias para esses atributos? Como esses atributos estão relacionados entre si? Qual o melhor modelo para predição de um bom vinho?

Essas perguntas podem ser respondidas não diretamente da mesma maneira com que foram formuladas, mas através da interpretação dos resultados obtidos nas operações aqui feitas.

É notável que a maior parte deste conjunto de dados é formado por amostras que não são vinhos classificados com notas 5 e 6. Uma pequena parte dos vinhos é classificada com nota 8, e uma parte menor ainda é classificada como nota 3. Portanto, parte deste trabalho era entender o que faz esses poucos vinhos se destacarem para atingirem a avaliação recebida.

Neste trabalho, foram considerados bons vinhos aqueles classificados com notas 7 e 8, o que totalizam 159 das 1143 amostras de vinho. A partir do heatmap apresentado na análise exploratória de dados foi possível visualizar que as variáveis mais correlacionadas com a qualidade do vinho eram o volume alcoólico e a acidez volátil. Pudemos ver que a maior parte dos vinhos estão entre volume alcoólico de 9-10% e volatilidade ácida entre 0.3-0.7g/dm^3. Os vinhos considerados bons costumam estar entre 10-13% de volume alcoólico e entre 0.25-0.45g/dm^3 de volatilidade ácida. Ainda assim, existem vinhos bons que fogem desta margem visualizada.

Foram escolhidos três métodos de aprendizado de máquina supervisionado, o SVM, KNN e RF. Ambos os três foram submetidos a diferentes formas de avaliá-los. A princípio, avaliamos os modelos fazendo uma divisão simples dos dados, onde 30% das amostras foram utilizadas para teste dos modelos que foram ajustados no volume restante de dados. Mesmo sem nenhum cuidado com os hiperparâmetros ou com técnicas de validação, pudemos obter bons resultados de predição, com valores de acurácia entre 0.8 e 0.9. Por outro lado, apenas o RF obteve um desempenho de destaque na métrica de AUC.

Em seguida, pudemos realizar a validação dos modelos através do método de validação cruzada. Uma vez que aproximadamente apenas 14% do nosso conjunto de dados são classificados como bons vinhos, podemos dizer que temos um conjunto de dados desbalanceados. Num conjunto de dados desbalanceado é importante que as proporções de amostras de ambas as categorias sejam mantidas em cada fold. Para isso, utilizamos o método de validação cruzada estratificada, o que garante a preservação dessas proporções. No geral, pudemos obter certamente resultados mais confiáveis sobre o desempenho dos modelos, mas não com valores muito distintos daqueles vistos anteriormente.

Sequencialmente foi avaliado o impacto da otimização de hiperparâmetros no desempenho dos modelos selecionados. O modelo KNN foi o que mais se beneficiou da otimização de hiperparâmetros, pois foi o modelo que mais teve ganho de desempenho. Em todas as três condições de avaliação, o modelo que apresentou melhor desempenho foi o RF, estando sempre com resultados de ~90% de acurácia.

Por fim, foi analisada a importância de cada atributo no processo de classificação dessas variáveis. A primeira maneira de investigar isso foi feita ajustando uma floresta aleatória e tomando os valores de \_feature importance\_ retornados pelo modelo. A distinção entre a importância dos definida pela RF quando utilizado o critério gini e entropia não pareceu ser significativa para este conjunto de dados, ainda assim pudemos notar que essa diferença existe e deve ser avaliada quando nos propomos a utilizar esse método de investigação de atributos. A outra maneira de avaliar a importância dos atributos foi utilizando a permutação dos valores dessas features. A partir da feature permutation pudemos avaliar o impacto de cada atributo nos três modelos aqui vistos. O modelo SVM apresentou pouco valor de importância para todos os atributos no conjunto, tendo apenas o \_total sulfur dioxide\_ ter algum destaque. Já nos modelos RF e KNN foi possível notar maior assertividade em definir quais as características mais relevantes no processo de classificação. Os atributos de \_total/free sulfur dioxide\_ parecem ter sido os mais relevantes para o modelo KNN, enquanto para o RF os destaques foram para \_alcohol, sulphates\_ e \_volatile acidity\_.

Quando comparamos os valores de importância pra cada atributo definido a partir da \_feature permutation\_ no modelo RF, podemos ver uma semelhança com o que foi visto no heatmap que apresenta a correlação entre as variáveis. Coincidentemente ou não, foi o RF quem apresentou o melhor desempenho geral e, em situação prática, o modelo a ser escolhido seria o RF com otimização de hiperparâmetros.

Mais estudos de seleção de atributo e a definição de mais hiperparâmetros seriam positivos para exploração deste problema.

# Referências bibliográficas

[1] Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decision support systems. 2009 Nov 1;47(4):547-53.

[2]