Projeto I

Guilherme Vieira & Natanael Sarmento

UFPB - Departamento de Estatística

Resumo

Frequentemente pesquisadores estão interessados em modelos classificatórios onde, através de variáveis explicativas, obtém-se uma resposta adequada para a situação, isto é, um modelo cuja precisão é alta. Com isso, é dirigido um estudo para um banco de dados com variáveis categóricas, referente a 23 espécies de cogumelos himênicos pertencentes à família Agaricus e Lepiota, o conjunto de dados foi reportado no Guia de campo da Audobon Society, e pode ser acessado pelo site do UCI - Machine Learning Repository. A ideia consiste em categorizar as espécies em comestível ou venenosa através de informações sobre o chapéu, estipe, anel micélio, entre outras. Neste escopo, compara-se as técnicas de arvore de decisão, Knn, Random Forest, maquina de vetores de suporte (SVM), regressão logística multinominal, Perceptron de neuronio unico, para propor o modelo final considerando a assertividade destes através do método validação cruzada (técnica k-folds e hold-out) .

I. Introdução

Frequentemente pesquisadores estão interessados em modelos classificatórios, onde, através de variáveis explicativas, obtém uma resposta adequada para a situação, isto é, um modelo cuja precisão é alta.

A Classificação é o processo de prever a classe de determinados pontos de dados. Às vezes, as classes são chamadas de destinos / marcadores ou categorias. A modelagem preditiva de classificação é a tarefa de aproximar uma função de mapeamento (f) de variáveis de entrada (X) para variáveis de saída discretas (y).

De uma maneira geral, estamos interessados em construir, com base em uma amostra de dados, uma regra de classificação e utilizá-la para classificar novos objetos.

Para a proposta de modelo, utiliza-se o conjunto de dados referente a 8124 registros das características físicas de 23 espécies de cogumelos himenicos das famílias Agaricus e Lepiota, juntamente com sua comestibilidade. A tarefa de classificação é determinar a comestibilidade, dadas as características físicas dos cogumelos.

Para tanto, opta-se pela construção de modelos classificatórios através dos métodos de arvore de decisão, Knn, Random Forest, maquina de vetores de suporte (SVM), regressão logística multinominal, Perceptron de neuronio unico, para propor o modelo final considerando a assertividade destes através do método validação cruzada .

II. Objetivos

Fixou-se a meta de propor modelos classificatórios para um conjunto de dados puramente categórico, afim de refinar e expandir os conhecimentos vistos em sala. Dentre os questionamentos levantados a respeito do bancos coletado, elencou-se duas duvidas:

O que caracteriza um cogumelo ser comestível ?

Quais informações dos cogumelos são realmente relevantes para que ocorra uma classificação satisfatória (um mínimo de 95% de assertividade) ?

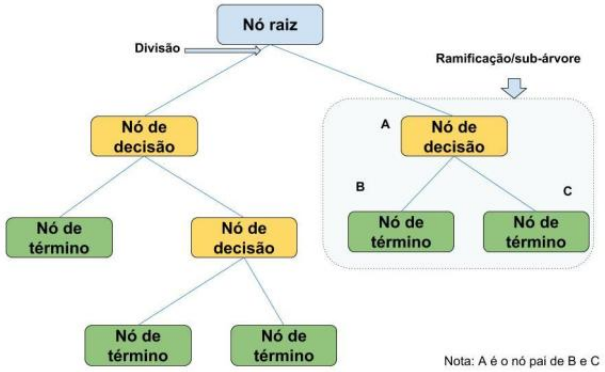
III. Sobre os dados

O banco de dados pode ser acessado através do link https://archive.ics.uci.edu/ml/datasets/mushroom. O conjunto de dados consiste em 8.124 observações com 23 colunas. Todo o conjunto de dados é uma entrada de um único caractere com significado único, podendo ser interpretados como fatores. A titulo de exemplo observe que a classificação recebe os acronimos ”e”e ”p”para comestivel e venenenoso respectivamente, e de maneira analoga seguem as variavies como, formato do chapéu para suas respectivas categorizações sino = b , conico = c , convexo = x, achatado = f, nodoso = k, afundado = s. Os valores ausentes no conjunto de dados cogumelo são identificados como ‘?’.

IV. Métodos

I. Árvore de Decisão

O aprendizado em árvore de decisão é uma das abordagens de modelagem preditiva usadas em estatística, mineração de dados e aprendizado de máquina. Ele usa uma árvore de decisão (como modelo preditivo) para passar de observações sobre um item (representado nos ramos) a conclusões sobre o valor alvo do item (representado nas folhas). Os modelos de árvore em que a variável de destino pode receber um conjunto discreto de valores são chamados de árvores de classificação;



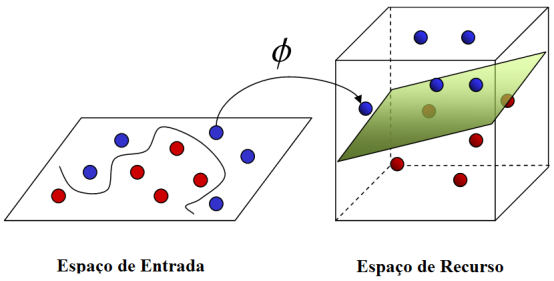
nessas estruturas em árvore, as folhas representam rótulos de classe e os ramos representam conjunções de recursos que levam a esses rótulos de classe. As árvores de decisão nas quais a variável de destino pode assumir valores contínuos (geralmente números reais) são chamadas de árvores de regressão.

Na análise de decisão, uma árvore de decisão pode ser usada para representar visual e explicitamente decisões e tomada de decisão. Os algoritmos para a construção de árvores de decisão geralmente funcionam de cima para baixo, escolhendo uma variável em cada etapa que melhor divide o conjunto de itens. Algoritmos diferentes usam métricas diferentes para medir o "melhor". Como estamos utilizando o software R para computação dos resultados, utiliza-se o pacote rpart dado que este contem um grande numero de métricas a serem utilizadas (como o indice de Gini e ganho de informação, para mais informações ver `https://cran.r-project.org/web/packages/rpart` . Eles geralmente medem a homogeneidade da variável de destino dentro dos subconjuntos. Essas métricas são aplicadas a cada subconjunto candidato e os valores resultantes são combinados (por exemplo, média) para fornecer uma medida da qualidade da divisão.

II. Vetor de suporte de Maquina (SVM)

O SVM (Support Vector Machines) é um método de classificação de dados que separa os dados usando hiperplanos. O conceito de SVM é muito intuitivo e facilmente compreensível. Se tivermos rotulado dados, o SVM poderá ser usado para gerar vários hiperplanos de separação, de modo que o espaço de dados seja dividido em segmentos e cada segmento contenha apenas um tipo de dados. A técnica SVM é geralmente útil para dados que não têm regularidade, o que significa dados cuja distribuição é desconhecida.

Na prática, o algoritmo de SVM consegue fazer ótimas classificações lineares, mas muitos problemas do mundo real não são linearmente separáveis. Para tanto, foi desenvolvido o método chamado de truque de Kernel, que consiste na transformação do conjunto de dados para outra dimensão a fim de separar os dados nessa outra dimensão e retornar posteriormente para a dimensão anterior com os dados já separados.



Este kernel é o que Alguns exemplos incluem gaussiano e radial. Portanto, o SVM também pode ser usado para dados não lineares e não requer nenhuma suposição sobre sua forma funcional. Também podemos interpretar os resultados produzidos pelo SVM através da visualização. Uma desvantagem comum do SVM está associada ao seu ajuste. O nível de precisão na previsão sobre os dados de treinamento deve ser definido. Em situações de negócios em que é necessário treinar o modelo e prever continuamente os dados de teste, o SVM pode cair na armadilha do ajuste excessivo. Esse é o motivo pelo qual o SVM precisa ser cuidadosamente modelado - caso contrário, a precisão do modelo pode não ser satisfatória. A técnica SVM está intimamente relacionada à técnica de regressão. Para dados lineares, podemos comparar o SVM com a regressão linear, enquanto o SVM não linear é comparável à regressão logística. À medida que os dados se tornam cada vez mais lineares, a regressão linear se torna cada vez mais precisa. No entanto, o ruído e o viés podem afetar gravemente a capacidade de regressão, nesses casos, o SVM é realmente útil. Para mais informações visitar

`https://cran.r-project.org/web/packages/e1071`

III. Regressão Logistica Multinomial

A regressão logística multinomial é a análise de regressão a ser conduzida quando a variável dependente é nominal com mais de dois níveis. Semelhante à regressão linear múltipla, a regressão multinominal é uma análise preditiva. A regressão multinominal é usada para explicar a relação entre uma variável dependente nominal e uma ou mais variáveis independentes.

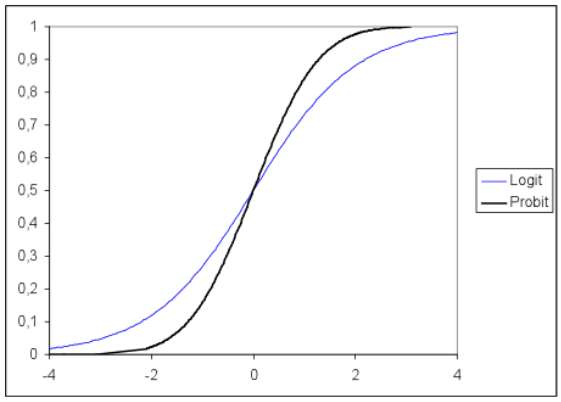
A regressão linear padrão requer que a variável dependente seja medida em uma escala contínua (intervalo ou razão). A regressão logística binária assume que a variável dependente é um evento estocástico. A variável dependente descreve o resultado desse evento estocástico com uma função de densidade (uma função de probabilidades acumuladas que variam de 0 a 1). Um ponto de corte (por exemplo, 0,5) pode ser usado para determinar qual resultado é previsto pelo modelo com base nos valores dos preditores.

Como podemos aplicar o princípio da regressão logística binária a uma variável multinomial (por exemplo, 1/2/3)?

Exemplo: Analisamos nossa classe de alunos que observamos por um período inteiro. No final do semestre, demos a cada aluno um jogo de computador como presente por seu esforço. Cada participante era livre para escolher entre três jogos - uma ação, um quebra-cabeça ou um jogo de esportes. Os pesquisadores querem saber como as notas dos alunos em matemática, leitura e escrita afetam sua escolha de jogo. Note que a escolha do jogo é uma variável dependente nominal com três níveis. Portanto, a regressão multinomial é uma abordagem analítica apropriada para a questão.

Como passamos da regressão logística binária para a regressão multinomial? A regressão multinomial é um modelo de multi-equação. Para uma variável dependente nominal com categorias k, o modelo de regressão multinomial estima equações k-1 logit.

O que são logits? A ideia básica por trás dos logits é usar uma função logarítmica para restringir os valores de probabilidade entre 0 e 1. Às vezes, um modelo de probit é usado em vez de um modelo de logit para regressão multinomial. O gráfico a seguir mostra a diferença entre um modelo logit e probit para valores diferentes. Ambos os modelos são comumente usados como a função de link na regressão ordinal. No entanto, a maioria dos modelos de regressão multinomial é baseada na função logit. Uma diferença notável entre funções geralmente é vista apenas em amostras pequenas, porque o probit assume uma distribuição normal da probabilidade do evento, enquanto o logit assume uma distribuição do log.



No centro da análise de regressão multinomial está a tarefa de estimar as probabilidades logarítmicas de cada categoria. No nosso exemplo de jogo de computador k = 3 com a última categoria como categoria de referência, a regressão multinomial estima funções de regressão k-1.

A regressão multinomial é semelhante à análise discriminante. A diferença prática está nas suposições de ambos os testes. Se as variáveis independentes são normalmente distribuídas, devemos usar a análise discriminante, porque é mais estatisticamente poderosa e eficaz.

IV. K-nn

Em estatística, o algoritmo de vizinhos k-mais próximos (k-NN) é um método não paramétrico proposto por Thomas Cover usado para classificação e regressão. Em ambos os casos, a entrada consiste nos k exemplos de treinamento mais próximos no espaço de caracteristicas em estudo. A saída depende se k-NN é usado para classificação ou regressão

k-NN é um tipo de aprendizado baseado em instância, ou aprendizado preguiçoso, onde a função é aproximada apenas localmente e todos os cálculos são adiados até a avaliação da função. Como esse algoritmo depende da distância para classificação, normalizar os dados de treinamento pode melhorar drasticamente sua precisão.

Tanto para classificação quanto para regressão, uma técnica útil pode ser atribuir pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuam mais para a média do que os mais distantes. Por exemplo, um esquema de ponderação comum consiste em dar a cada vizinho um peso de 1 / d, onde d é a distância ao vizinho.

Os vizinhos são obtidos de um conjunto de objetos para os quais a classe (para classificação k-NN) ou o valor da propriedade do objeto (para regressão k-NN) é conhe-

cido. Isso pode ser considerado o conjunto de treinamento para o algoritmo, embora nenhuma etapa de treinamento explícita seja necessária.

Uma peculiaridade do algoritmo k-NN é que ele é sensível à estrutura local dos dados.

V. Random Forest

Florestas aleatórias ou florestas de decisão aleatória são um método de aprendizagem de conjunto para classificação, regressão e outras tarefas que operam construindo uma infinidade de árvores de decisão no momento do treinamento e gerando a classe que é o modo das classes (classificação) ou predição média / média (regressão) das árvores individuais.

As florestas de decisão aleatória corrigem o hábito das árvores de decisão de se ajustar ao seu conjunto de treinamento. Geralmente superam as árvores de decisão, mas sua precisão é menor do que as árvores com aumento de gradiente. No entanto, as características dos dados podem afetar seu desempenho.

O primeiro algoritmo para florestas de decisão aleatória foi criado por Tin Kam Ho usando o método do subespaço aleatório, que, na formulação de Ho, é uma forma de implementar a abordagem de ”discriminação estocástica”para classificação proposta por Eugene Kleinberg.

Uma extensão do algoritmo foi desenvolvida por Leo Breiman e Adele Cutler, registraram ”Random Forests”como uma marca comercial (em 2019, de propriedade da Minitab, Inc.). A extensão combina a ideia de ”ensacamento”de Breiman e a seleção aleatória de recursos, a fim de construir uma coleção de árvores de decisão com variância controlada.

Florestas aleatórias são frequentemente usadas como modelos de ”caixa preta”em empresas, pois geram previsões razoáveis em uma ampla gama de dados, embora exijam pouca configuração em pacotes como o scikit-learn.

VI. Perceptron - 1 neuronio

No aprendizado de máquina, o perceptron é um algoritmo de aprendizado supervisionado de classificadores binários. Um classificador binário é uma função que pode decidir se uma entrada, representada por um vetor de números, pertence ou não a alguma classe específica. É um tipo de classificador linear, ou seja, um algoritmo de classificação que faz suas previsões com base em uma função de preditor linear combinando um conjunto de pesos com o vetor de características.

VII. validação cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Busca-se então estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, o uso de alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento), sendo os subconjuntos restantes (dados de validação ou de teste) empregados na validação do modelo.

O método de validação cruzada denominado k-fold consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros, fazendo-se o cálculo da acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. A figura abaixo mostra o esquema realizado pelo k-fold.

Ao final das k iterações calcula-se a acurácia sobre os erros encontrados, através da equação descrita anteriormente, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

No R, para modelos de classificação, as funções sensibilidade, especificidade, posPredValue e negPredValue podem ser usados para caracterizar o desempenho do modelo onde houver duas classes. Então quanto mais proximo de 1 melhor sera o modelo

VIII. Do tratamento dos dados

Por se tratar de um conjunto inteiramente categorico, duas coisas são necessarias, fundamentalmente para que os modelos possam ser construidos adequadamente. A primeira é a transformação das variaveis em fatores e a segunda é a transformação desses fatores em valores inteiros, para que ocorra uma campatibilidade entre os dados e as exigencias para confecção dos modelos de aprendizagem de maquina. Se fez necessaria a renomeação das variavies, para ter uma apresentação mais agradavél, assim como estabelecer como semente ”seed(2020)”para replicações do estudo.

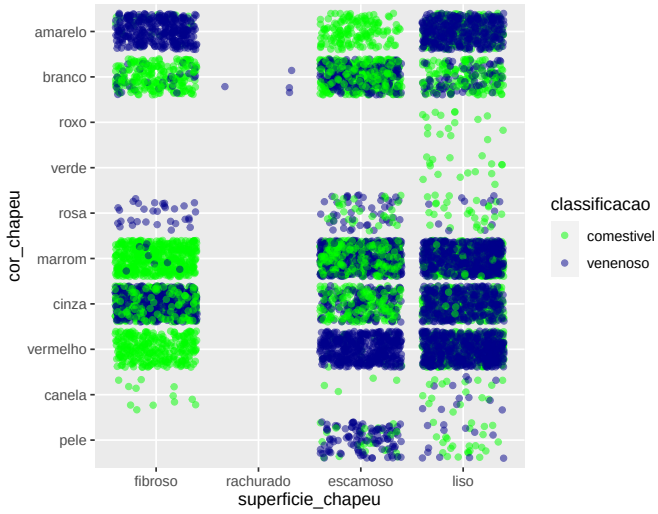
Além disso apenas uma variável (stalk-root) parece conter valores ausentes(2480 ausencias). Devido a um

grande número de valores ausentes em stalk-root, esse recurso foi removido.

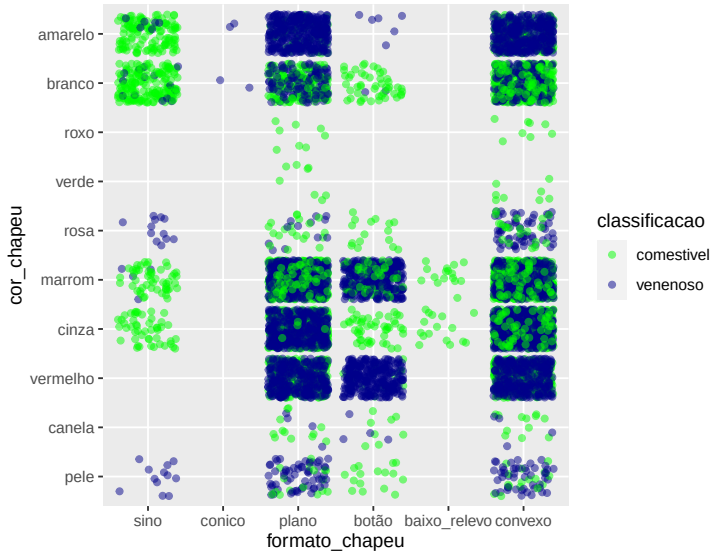
V. Resultados e Discussão

I. Analise Descritiva

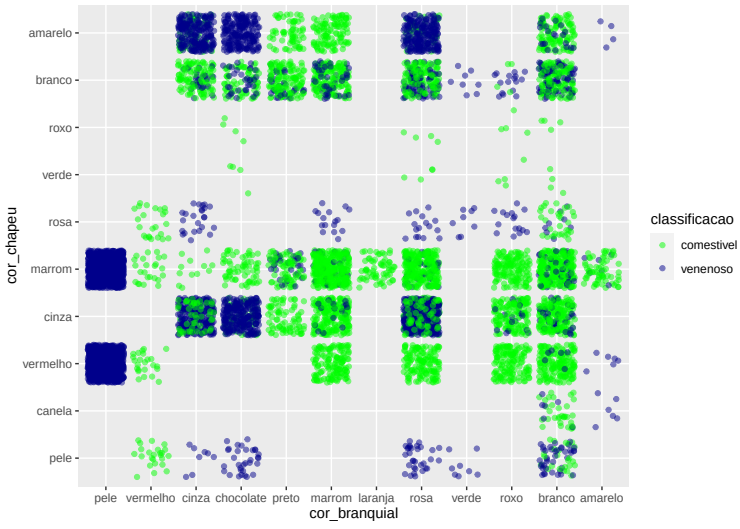
Mediante a analise descritiva temos uma clareza dos dados, ao ponto que temos um norte de quais variáveis podem ser significativas para os modelos que venhamos a esboçar. Inicialmente temos que 52% da amostra é classificada como comestível, e 48% é constituída dos cogumelos venenosos. Com a ajuda da ilustração, a relação entre superfície e cor chapéu discrimina a classificação.



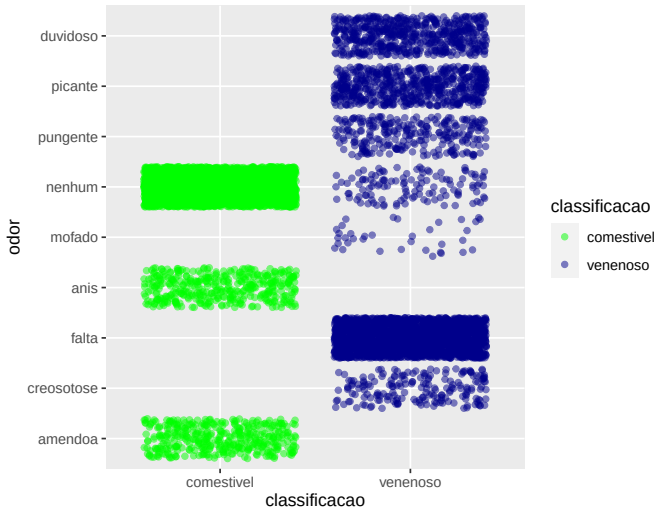
A superfície fibrosa nos mostrou mais observações classificadas como comestível, enquanto que a lisa mostra o oposto, salvo o caso de quando a cor do chapéu é verde ou roxa. A superfície escamosa com a cor do chapéu amarela também nos mostra total segurança para consumo, enquanto que com o chapéu vermelho totalmente venenoso. Enquanto na ilustração seguinte, O formato de sino se mostra mais seguro para consumo em relação aos demais formatos, com exceção do no formato de baixo relevo sinaliza da mesma forma, contudo temos poucas observações do mesmo.



Vemos a seguir que para um cogumelo com cor do chapéu vermelho ele é comestível quando a cor do branquio é vermelho, marrom, rosa e roxo. Ou quando a cor do branquio é vermelho essa conclusão se dá para quando a cor do chapéu é cor de pele, vermelha, marrom, e rosa Quando a cor branquial for cor de pele temos que o cogumelo é venenoso.

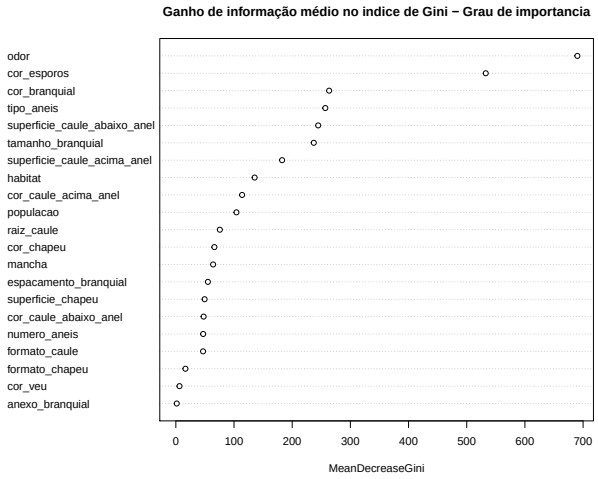


O odor é definitivamente um preditor informativo. Basicamente, se cheira é duvidoso, picante ou pungente, fique longe. Se cheira a anis ou amêndoa, você pode ir em frente. Se não cheira nada, você tem mais chances de ser comestível do que não.

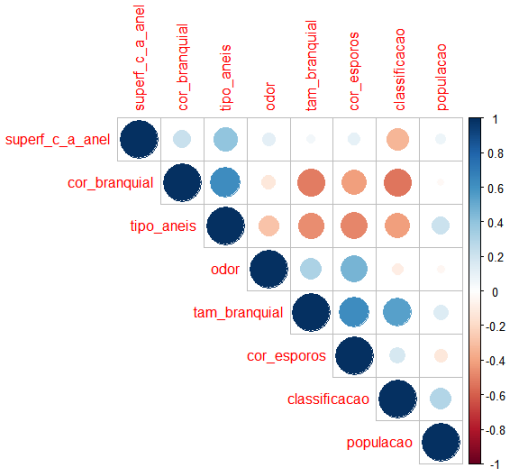


II. redução de dimensionalidade e correlações

Há um punhado de tecnicas utilizadas comumente para a redução de dimensionalidade de grandes conjuntos de dados. Considerando a especifidade dos dados (completamente categoricos), fica-se um tanto dificil a aplicação da analise de componentes principais. Dessa forma, recorre-se ao Random Forests / Ensemble Trees, para elencar a as principais variaveis, considerando o ganho de informação. A ilustração a seguir é a saída da aplicação dessa tecnica. Observe que odor é considerada a variavel mais importante, seguida de cor dos esporos e cor das branquias. Estabelecendo um corte de 145, onde as variaveis com importancia maior que serão consideradas nos estudos dos modelos adiante, assim conseguindo fazer uma redução de dimensionalidade onde o banco com 21 variáveis exceto a variavel resposta, vemos que 7 variaveis tem grau de importancia alto.



Após a redução de dimensionalidade, vamos fazer uma breve analise das correlações entre essas variaveis. O plot a seguir, consegue sumarizar essa informação



Observa-se uma correlação positiva proxima de 0.8 entre as cor branquial e tipo de anel, entre odor e cor dos esporos, entre tamanho branquial e cor dos esporos, e entre tamanho branquial e a classificação.

Por sua vez, obseva-se uma correlação negativa proxima de 0.8 entre cor branquial e tamanho branquial, entre cor branquial e classificação.

VI. Dos Modelos

Note que a variavel resposta é categorica e os banco de treinamento e de teste que obtemos na validação cruzada precisa ter a mesma proporção, onde no banco original a variavel ”classificação”apresenta 0.52 comestivel e 0.48 venenosso e os bancos de teste e treinamento essa proporções precisam ser mantidas.

Como estamos lidando com uma tarefa de classificação, na validação cruzada para k-fold, com k e cv iguais a 10, calculamos a matriz de confussão e obtemos o seguintes indicadores: Sensitivity, Specificity, Prevalence, PPV(Pos Pred Value), NPV(Neg Pred Value), Detection Rate, Detection Prevalence, Balanced Accuracy, Precision , Recall e F1.

A seguir a tabela com esses criterios, com 19 das variaveis, pois o knn não apresentou um bom desempenho, e em seguinda uma mesma tabela com apenas as 7 variaveis referentes a redução de dimensionalidade

	Arvore	Random	SVM	Knn	Logist.Mult	Perceptron
Sensitivity	1.00	1.00	1.00	1.00	1.00	0.997
Specificity	0.988	1.00	0.996	0.999	1.00	0.951
Pos Pred Value	0.989	1.00	0.996	0.999	1.00	0.960
Neg Pred Value	1.00	1.00	1.00	1.00	1.00	0.997
Precision	0.989	1.00	0.996	0.999	1.00	0.960
Recall	1.000	1.00	1.00	1.00	1.00	0.997
F1	0.994	1.00	0.998	0.999	1.00	0.977
Prevalence	0.518	0.518	0.518	0.518	0.518	0.518
Detection Rate	0.518	0.518	0.518	0.518	0.518	0.516
Detection Prevalence	0.523	0.518	0.520	0.518	0.518	0.540
Balanced Accuracy	0.994	1.00	0.998	0.999	1.00	0.974

	Arvore	Random	SVM	Knn	Logist.Mult	Perceptron
Sensitivity	1.00	1.00	1.00	NA	1.00	0.997
Specificity	0.988	1.00	0.988	NA	1.00	0.958
Pos Pred Value	0.989	1.00	0.989	NA	1.00	0.966
Neg Pred Value	1.00	1.00	1.00	NA	1.00	0.997
Precision	0.989	1.00	0.989	NA	1.00	0.966
Recall	1.00	1.00	1.00	NA	1.00	0.997
F1	0.994	1.00	0.994	NA	1.00	0.980
Prevalence	0.518	0.518	0.518	NA	0.518	0.518
Detection Rate	0.518	0.518	0.518	NA	0.518	0.516
Detection Prevalence	0.5285	0.518	0.524	NA	0.518	0.536
Balanced Accuracy	0.993	1.00	0.994	NA	1.00	0.978

Através do metodo hold-out (80% treinamento e 20% teste) temos a tabela a seguir, com os criterios de validação:

	Arvore	Random	SVM	Knn	Logist.Mult	Perceptron
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	0.9923	1.00	0.997	1.00	1.00	0.973
Pos Pred Value	0.9929	1.00	0.9997	1.00	1.00	0.976
Neg Pred Value	1.00	1.00	1.00	1.00	1.00	1.00
Prevalence	0.518	0.518	0.518	0.518	0.518	0.518
Detection Rate	0.518	0.518	0.518	0.518	0.518	0.518
Detection Prevalence	0.5216	0.518	0.519	0.518	0.518	0.530
Balanced Accuracy	0.996	1.00	0.999	1.00	1.00	0.9866
Kappa	0.992	1.00	0.997	1.00	1.00	0.9741
Accuracy	0.996	1.00	0.9988	1.00	1.00	0.9871

No geral, todos os modelos se sairam muito bem. Contudo, partindo de um lógica que um individuo utilize tais tecnicas para verificar a comestibilidade de um fungo, é interessante que a acuracia deste seja perfeita, pois o erro minimo pode resultar na morte daquele que o consumir. Nesse contexto, chama-se a atenção para os modelos de Random Florest, Logist Multinomial, e o Knn, contudo este ultimo não se torna interessante uma vez que uma quantidade maior de informação é necessaria para a classificação.

VII. Conclusão

O conjunto de dados de cogumelos é analisado de quatro maneiras. O primeiro envolveu o uso de gráficos para explorar a contribuição dos atributos na decisão da comestibilidade do cogumelo. Os plots fazem parte da análise bi-variável, onde não existe um único atributo que possa servir suficientemente como fator decisivo. Alguns atributos não desempenham absolutamente nenhum papel na tomada de decisões e, portanto, podem ser completamente ignorados. Este é um exemplo de redução de dimensionalidade para obter melhor eficiência. Para as partes em diante foi necessário dividir os dados em conjuntos de treinamento, e teste, para que o modelo preditivo obtenha observações completamente novas para trabalhar. Por sua vez, é necessário avaliar a precisão do modelo atavrés das tecnicas de validação cruzada (k-folds e hold-out).

É usado o "randomforest"apenas rapidamente e o coeficiente de Gini para calcular a importância dos atributos em decidir a classe à qual um determinado cogumelo pertence. A importância das variáveis é calculada, plotada e exibida em ordem decrescente. Depois de entender a importância das variáveis, é construido um plot das correlações no conjunto de dimensionalidade reduzida. É elaborada uma árvore de decisão que ajude a decidir a comestibilidade do cogumelo (verificar o arquivo Markdown).

Para as terceira e quarta parte são utilizados os mecanismos no software para gerar modelos de alimentação continua, como no caso o SVM, perceptron, e o Logistico multinomial, e suas respectivas validações. Este ultimo se mostrou com acuracia perfeita.

Referências

[1] R Core Team, R Foundation for Statistical Computing, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, https://www.R-project.org/, 2018