Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Ricardo Iwagoro Piekarz Umeyama

Análise preditiva em testes de estabilidade dentro do desenvolvimento de produtos no setor de cosméticos

Curitiba
2020

Ricardo Iwagoro Piekarz Umeyama

Análise preditiva em testes de estabilidade dentro do desenvolvimento de produtos no setor de cosméticos

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Walmes Marques Zeviani

Curitiba

2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Análise preditiva em testes de estabilidade dentro do desenvolvimento de produtos no setor de cosméticos

Ricardo Iwagoro Piekarz Umeyama[1]
Walmes Marques Zeviani[3]
Wagner Bonat[3]

Resumo

A realização de teste de estabilidade é uma etapa essencial no desenvolvimento e comercialização de produtos. É responsabilidade da empresa a seguridade do teste e cabe a Anvisa a fiscalização. Cada componente, variáveis extrínsecas ou intrínsecas podem afetar a estabilidade de um produto. Os testes seguem um protocolo e recomendações conforme o Guia de Estabilidade de Produtos Cosméticos - Anvisa [1] e devem seguir determinadas condições de armazenamento, parâmetros de avaliação e periodicidade. O objetivo deste trabalho é encontrar uma metodologia de aprendizado de máquina que possibilite predizer com um bom nível de acurácia se o estudo sofrerá alteração ou não ao final da avaliação t90. O presente estudo teve início com a criação de um banco de dados relacional, agrupando dados de duas tabelas. Foram aplicados os modelos de Naive Bayes, Regressão logística com regularização(GLM) e Support vector machine. Como métrica de avaliação foi utilizada a matriz de confusão, acurácia e especificidade. Foram gerados 5 dataframes conforme a FAMILIA que o estudo pertence e suas características. O modelo com melhor desempenho foi o GLM, utilizando *kfold = 10* com uma acurácia média entre as FAMILIAS de 92%.
Palavras-chave: ANVISA, CRISP-DM, classificação supervisionada, caret, GLM, Naive Bayes, SVM, matriz de confusão, acurácia, especificidade, varImp, linguagem R.

Abstract

Performing a stability testing has been an essential phase in the development and selling of products. The company is responsible for the suitability of test, and the inspection belongs to Anvisa (Brazilian Health Regulatory Agency) [1]. On each component, extrinsic and intrinsic variables may be affecting the stability of a product. Product testing follows the guidance of "Guia de Estabilidade de Produtos Cosméticos - Anvisa"and must follow recommendations for storage conditions, test parameters and its frequency. This final paper intends to find a machine learning methodology, which may help predicting with a good accuracy level whether the study suffers modification or not at the end of t90 estimation. This study has begun on a relational database with grouped data of two tables. Naive Bayes Classifiers, Logistic regression with regularization(GLM) and Support vector machine had been applied to. Confusion matrix, accuracy and specificity had been used as evaluation metrics. This had generated five data frames for the clusters and characteristics which the study belongs to. GLM has had the best performance with kfold = 10 *and 92% of average accuracy among clusters.*
***Keywords**: ANVISA, CRISP-DM, supervised classification, caret, GLM, Naive Bayes, SVM, confusion matrix, accuracy, specificity, varImp, R Language.*

1 Introdução

O estudo de estabilidade[2] tem por objetivo avaliar a capacidade de um produto manter as características organolépticas, físico-químicas, microbiológicas, segurança e eficácia. Assim, o estudo da estabilidade deve ser visto como um requisito necessário para a garantia da qualidade do produto e não somente como uma exigência do órgão regulamentador. Cabe a empresa a responsabilidade de avaliar a estabilidade de seus produtos antes de disponibilizá-los ao consumo. No Brasil, é de responsabilidade da Agência Nacional de Vigilância Sanitária – Anvisa, regulamentar, fiscalizar, controlar a produção e a comercialização de produtos cosméticos, para propiciar produtos seguros e com qualidade no mercado. Produtos expostos ao consumo e que apresentem problemas de estabilidade, além de descumprirem os requisitos técnicos de qualidade, podem ainda colocar em risco a saúde do consumidor configurando infração sanitária. A apresentação dos dados de estabilidade é exigida no ato da regularização do produto pela Anvisa.

O estudo da estabilidade de produtos cosméticos fornece informações que indicam o grau de estabilidade de um produto nas variadas condições a que possa estar sujeito desde sua fabricação até o término de sua validade. Essa estabilidade varia com o tempo e em função de fatores que aceleram ou retardam alterações nas características ou propriedades. Modificações dentro de limites determinados podem não configurar motivo

[1]Aluno do programa de Especialização em Data Science & Big Data, riwagoro@gmail.com.
[3]Professor do Departamento de Estatística - DEST/UFPR.

para reprovar o produto. O estudo da estabilidade de produtos cosméticos contribui para orientar o desenvolvimento e aperfeiçoamento da formulação, estimar o prazo de validade e auxiliar no monitoramento da estabilidade, produzindo informações sobre confiabilidade e segurança dos produtos.

Deve-se realizar esse estudo, durante o desenvolvimento de novas formulações e de lotes-piloto de laboratório e de fábrica, quando ocorrerem mudanças significativas no processo de fabricação para validar novos equipamentos ou processo produtivo, quando houver mudanças significativas nas matérias-primas do produto ou quando ocorrer mudança significativa no material de acondicionamento que entra em contato com o produto.

Cada componente, ativo ou não, pode afetar a estabilidade de um produto. Variáveis relacionadas à formulação, ao processo de fabricação, ao material de acondicionamento e às condições ambientais e de transporte podem influenciar na estabilidade do produto. Conforme a origem, as alterações podem ser classificadas como extrínsecas, quando determinadas por fatores externos (tempo, temperatura, material de luz/oxigênio, umidade, material de acondicionamento, microrganismos e vibração); ou intrínsecas (incompatibilidade física ou química) quando determinadas por fatores inerentes à formulação.

Os aspectos de análises considerados na estabilidade a avaliar são:

- Físicos: devem ser conservadas as propriedades físicas originais como aspecto, cor, odor, uniformidade;
- Químicos: devem ser mantidos dentro dos limites especificados a integridade da estrutura química, o teor de ingredientes, valor de pH, viscosidade, densidade e em alguns casos, o monitoramento de ingredientes da formulação;
- Microbiológicos: devem ser conservadas as características microbiológicas, conforme os requisitos especificados;
- Funcionalidade: os atributos do produto devem ser mantidos sem alterações quanto ao efeito inicial proposto;
- Segurança: não devem ocorrer alterações significativas que influenciem na segurança de uso do produto.

As condições de armazenagem mais comuns são 5 : temperatura ambiente 25ºC, elevada 40ºC, baixa 5ºC, exposição à luz fluorescente em ambiente 25ºC e led em ambiente 25ºC. Os produtos devem ser armazenados em mais de uma condição de armazenagem para que se possa avaliar seu comportamento nos diversos ambientes a que possa ser submetido, simulando as características da zona climática onde os produtos serão produzidos e ou comercializados.

A periodicidade da avaliação das amostras é que seja avaliada inicialmente no tempo zero, 24 horas e aos 7º, 15º, 30º, 60º e 90º dias após o início do armazenamento em cada condição e aspectos de análises.

A interpretação dos dados obtidos durante o estudo da estabilidade é avaliada em comparação à amostra-padrão que é registrada no tempo zero. Geralmente define-se limites de aceitação para as características avaliadas e a amostra-padrão deverá permanecer inalterada durante toda a vida útil do produto. A amostra pode ser classificada segundo os seguintes critérios:

- Normal (sem alteração): sem impacto;
- Levemente modificado: alteração dentro do limite de aceitação;
- Modificado: reprovação do produto.

O objetivo deste trabalho é encontrar uma metodologia de aprendizado de máquina que possibilite predizer se o produto sofrerá alteração ou não ao final da avaliação que ocorre no tempo do 90º dia. Pretende-se, com a conclusão deste trabalho e implementação desse novo processo, a redução do tempo de análise de estabilidade para determinados tipos de estudos e com isso acelerar o desenvolvimento de novos produtos.

2 Materiais e Métodos

Para realização deste estudo considerou-se a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) que é uma metodologia de mineração de dados, que direcionam a descoberta do conhecimento para tomada de decisão sobre dados em grande volume. Segundo Chapman [3], a metodologia CRISP-DM é composta por 6 fases organizadas de maneira cíclica, cujo fluxo é não unidirecional, possibilitando ir e voltar entre as suas fases e tarefas. As seções a seguir demonstram a aplicação do modelo.

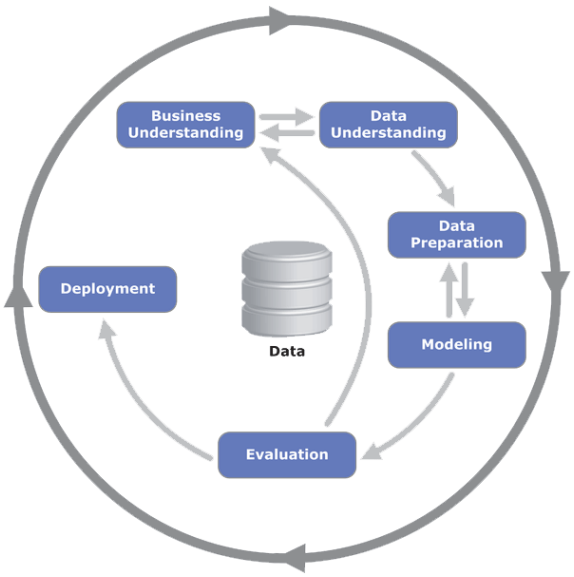


Figura 1: CRISP-DM diagrama do processo Fonte: Kenneth Jensen / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0).

2.1 Entendimento do Negócio

O objetivo deste trabalho é ter um método para rapidamente prever produtos com suspeita de deterioração baseado na composição e com isso poder acelerar o tempo de desenvolvimento de produtos, diminuindo o risco de casos que apresentam deterioração, com economia de matéria-prima, tempo e sem risco para o consumidor.

2.2 Entendimento dos Dados

2.2.1 Coletando dados iniciais

Os dados utilizados foram cedidos por uma empresa de cosméticos no mercado brasileiro, sob condição de manutenção de sigilo de todas as informações sensíveis e completa anonimização dos nomes das variáveis. Foram extraídos do software SAP de duas tabelas distintas por conexão ODB com o software R.

2.2.2 Descrevendo os dados

A tabela 1 corresponde a uma amostra de 54 mil estudos no período de Janeiro de 2017 a Agosto de 2019, com 11 variáveis categóricas que representam dados do número da formulação, categoria, família, aspectos, condições de armazenamento e os resultados das amostras no tempo 90º dia. Esses resultados estão categorizados da seguinte forma:

- Normal (sem alteração): 0;
- Levemente modificado: 0;
- Modificado: 1.

Tabela 1: Dicionário do conjunto de dados.

Aspecto	Conteúdo
*t*90	(binária) Houve alteração no *t*90 dias.
E.g.	Normal (0), Modificado (1).
Formulação	(nominal) Código da análise.
E.g.	1, 2, 3 etc.
Família	(nominal) Família a que pertence o produto.
E.g.	(nominal) Cabelo, Facial, Maquiagem Rosto etc.
Análise	(nominal) Aspectos analisados no estudo.
E.g.	(nominal) Cor Formulação, Odor Formulação, pH, Viscosidade etc.
Aspecto Análise	(nominal) Condição de armazenagem do produto.
E.g.	(nominal) AMB/25, EST/40, GEL/5, LUZ FLU e LUZ LED.

A tabela 2 traz 522 mil registros referente a formulação, componentes, percentual de concentração para cada formulação, representando um total de 5086 variáveis.

Tabela 2: Dicionário do conjunto de dados.

Aspecto	Conteúdo
Formulação	(nominal) Código da análise.
E.g.	1, 2, 3 etc.
Componente	(nominal) Componentes que pertencem ao código da análise.
E.g.	(nominal) bPxD, NndU, AaYY etc.
TT	(numérica) Percentual de concentração do componente da análise.
E.g.	(numérica) 0 < percentual =< 1.

2.2.3 Explorando os dados

Analisando a tabela 1, temos 17% dos registros que chegam no t90 com resultado igual a 1 (Modificado), porém temos um percentual bem elevado, com um total de 83% dos registros que durante todo o período de análise não sofre nenhum tipo de alteração que fuja do limite tolerável, sendo classificado como 0 (normal) ou (levemente modificado), caracterizando esses dados em um dataset desbalanceado, dado que nosso interesse é na classe minoritária, ou seja, determinar se o resultado irá modificar no t90.

Tabela 3: Resultado da tabela 1 referente a análise no tempo t90 por tipo de resultado.

t90	Quantidade	Percentual
0	45056	83%
1	9245	17%

2.2.4 Verificando a qualidade dos dados

Os dados extraídos possuem algumas inconsistências de valores, dados incompletos e faltantes, causados por perda de informação dentro do sistema, erro de digitação, campos sem regra de negócio associada, além de estudos que foram cancelados por algum motivo. Esses dados foram tratados em um processo de ETL (extract, transform, load) e removidos da tabela 1, onde houve uma perda de 15% nos registros.

2.3 Preparação dos Dados

A preparação dos dados é de fundamental importância para o aprendizado do algoritmo. É preciso garantir que estamos selecionando todos os atributos relevantes e que nossa amostra consiga representar o real problema de negócio.

2.3.1 Selecionar dados

A tabela 1 possui variáveis categóricas e suas características, que representam a família, aspectos, condições de armazenamento e os resultados no tempo 90º dia. Além

disso, contém dados de formulação, seus componentes e % de concentração que constam na tabela 2.

2.3.2 Dados Limpos

Foram desconsiderados da tabela 1 os registros onde não existia a leitura da informação no t90. Essa falta de informação no t90 ocorre devido a um projeto cancelado ou uma análise que foi cancelada por um determinado motivo que não conseguimos rastrear.

2.3.3 Construir Dados

Dos 19 subconjuntos do atributo FAMILIA, foram utilizados as 5 principais FAMILIAS considerando a relevância para a área de negócio, atributos e tamanho das bases. As variáveis categóricas ANALISE e ASPECTO ANALISE foram transformadas em variáveis binárias onde o número 1 representa o valor afirmativo e o 0 negativo. Para o atributo ASPECTO_ANALISE da tabela 1 o registro da informação no t90 se dá de duas formas: para características como aspecto formulação, aplicação formulação, cor formulação, odor formulação os dados registrados são representados em: normal, levemente modificado e modificado. Para as características Ph e Viscosidade o registro é feito com a leitura do valor real no tempo de análise. Essa leitura é convertida em % de variação entre os tempos conforme abaixo:

- Normal ou levemente modificado: 0% a 9,99%;
- Modificado: maior que 10%.

2.3.4 Integrar dados

A criação da tabela 3, foi feita integrando a tabela 1 e tabela 2, para isso usamos como registro chave o atributo FORMULACAO existente nas duas tabelas. Foi utilizado o software R e o pacote `tidyverse` [4].

2.3.5 Formatar Dados

Com a tabela 3 criada definindo X como um vetor de características da amostra,

$$X = [x_1, x_2, ..., x_p].$$

E como vetor de rótulos t90,

$$\text{t90}= \begin{cases} 0, & \text{normal ou levemente modificado,} \\ 1, & \text{modificado.} \end{cases}$$

2.4 Modelagem

As características determinísticas de cada estudo estão associadas ao grupo de FAMILIA que a pertence, com isso foi dividida a tabela 3 de acordo com a sua FAMILIA, dando um total de 5 dataframes. Para cada dataframe foi aplicado validação cruzada, técnica utilizada em problemas de aprendizado de máquina que consiste em particionar os dados de modo que o método seja treinado usando uma parte e avaliando o restante. Isso permite observar o desempenho do método com dados novos do ponto de vista de qualidade das predições. A partição de dados foi realizada da seguinte forma: treino 80% e teste 20%. Para os três modelos propostos foram usados as mesmas bases de treino e teste além do mesmo *kfold=10* para avaliação, com otimização dos hiperparâmentos. Os modelos [5] selecionados são modelos interpretáveis, com baixo tempo de processamento e estão listados abaixo.

- Naive Bayes(NB) - Classificador probabilístico baseado no teorema de Bayes e utilizada uma forte premissa de independência condicional;
- Regressão logística com regularização(GLM) - Tem como objetivo produzir, a partir do conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias;
- Support vector machine(SVM) - É uma técnica de classificação e tem por objetivo encontrar uma linha de separação chamada de hiperplano entre um conjunto de exemplos. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma dos conjuntos.

Além dos modelos acima, foram testados outros modelos [5] como Random forest e Extreme gradient boosting, mas devido ao alto tempo computacional e a necessidade de algoritmos com um tempo computacional mais rápido, eles foram descartados para esse estudo.

Para execução dos modelos propostos foi escolhido o software R juntamente com o pacote `caret` (3).

2.5 Avaliação

Avaliar o desempenho dos métodos de classificação exige a escolha de uma ou mais medidas que possibilitem a correta leitura dos resultados e consequentemente a escolha do melhor método. Em geral, a medida mais comum é a acurácia, que apenas conta o número de acertos (classificações corretas) na amostra teste. Porém, para dados desbalanceados essa medida pode não fornecer o real desempenho do classificador. Essa contagem pode ser resumida em uma tabela denominada matriz de confusão [6]. Queremos que o modelo tenha uma boa predição em classificar estudos que irão modificar (Modificado). Será utilizado a classe 1 (y=1) para as medidas de avaliação.

- TP: Taxa verdadeiros positivos (normal e levemente modificado);
- FN: Taxa de falsos negativos (a classe verdadeira é negativo, mas a saída do modelo é positivo - errou);
- FP: Taxa de falsos positivos (a classe verdadeira é positivo, mas a saída do modelo é negativo - errou);
- TN: Taxa de verdadeiros negativos (modificado).

Tabela 4: Matriz de confusão.

		Real	
		y=0	y=1
Predito	y=0	TP	FP
	y=1	FN	TN

A partir da Tabela 4 algumas medidas podem ser calculadas. Neste estudo serão avaliadas com objetivo de minimizar o erro de classificação *(FP)*, dado que a predição do estudo y=0 quando o verdadeiro rótulo é y=1 representando que o estudo irá modificar. São elas:

A *acurácia* dá a proporção de predições corretas.

$$\text{Acurácia : A}=\frac{TP+TN}{TP+FP+TN+FN}.$$

A *sensibilidade* mede a força do modelo prever um resultado positivo.

$$\text{Sensibilidade : R } = \frac{TP}{TP+FN}.$$

A *especificidade* mede a força do modelo prever um resultado negativo.

$$\text{Especificidade : E} = \frac{TN}{TN+FP}.$$

A *exatidão* mede a precisão de um resultado previsto como positivo.

$$\text{Exatidão : P} = \frac{TP}{TP+FP}.$$

O *F1* é a média harmônica entre precisão e a sensibilidade.

$$\text{F1 : F} = 2 \cdot \frac{P \cdot R}{P+R}.$$

2.6 Implementação do Modelo

A implementação coloca o modelo em produção para que possa ser utilizado. Ele coloca fim ao estudo, mas é necessário garantir o monitoramento dos resultados e adaptar o modelo sempre que necessário.

3 Resultados e Discussões

A matriz de confusão foi gerada para os três modelos propostos e os 5 subconjuntos da variável FAMILIA. Conseguimos então extrair as métricas abaixo para avaliar os modelos.

3.1 Acurácia

A acurácia entre a base treino e base teste de cada metodologia aplicada para os 5 subconjuntos do atributo FAMILIA estão contidas na Figura 2, para auxiliar na escolha de qual modelo tem melhor desempenho. O algoritmo NB teve um pior desempenho entre os modelos, com uma acurácia de 84% contra 91% do SVM e 92% do GLM. Entre as 5 FAMILIAS os modelos SVM e GLM mantiveram performances semelhantes, sendo o modelo GLM levemente superior. A acurácia é a métrica mais comum para problemas de classificação, embora nem sempre seja adequada, sendo potencialmente perigosa quando usada em respostas categóricas desbalanceados. Para tal, podemos usar métricas alternativas como a especificidade que permitem que o desempenho do modelo seja considerado com foco na classe minoritária, neste caso, classe: Modificado (y=1)).

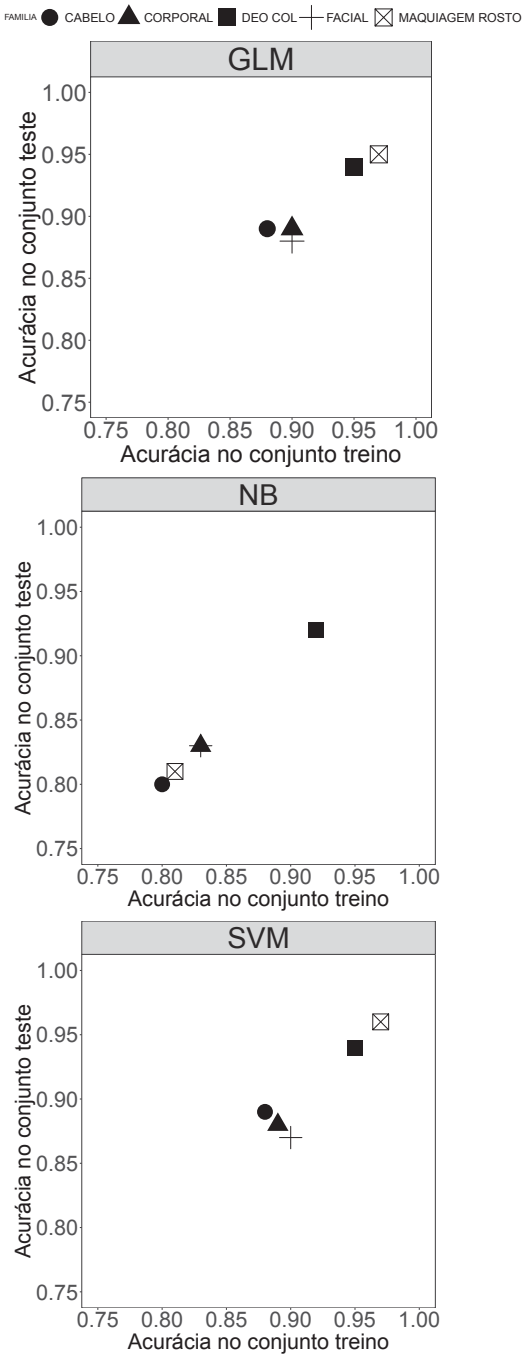


Figura 2: Acurácia nos conjuntos de treino e teste para cada modelo e cada FAMILIA.

3.2 Sensibilidade

A sensibilidade mede a força do modelo em prever um resultado positivo (TP). Porém essa medida tem o foco na classe majoritária, neste caso, classe normal ou levemente modificado (y=0)) que não é nossa classe de interesse no estudo. Os modelos testados obtiveram um boa sensibilidade sendo capaz de predizer nossa classe *y=0*. Para o modelo NB tivemos um resultado de 100% e quando analisamos, torna-se um erro de predição dado que ele classificou todo conjunto em *y=0* sendo que em nosso conjunto de dados de treino e teste possui valores *y=0* e *y=1*, podendo, desta forma, descartar esse modelo para esse estudo.

Tabela 5: R: Sensibilidade para os modelos propostos e seus subconjuntos.

	Modelo		
	GLM	NB	SVM
CABELO	0.94	1.00	0.94
CORPORAL	0.96	1.00	0.94
DEO COL	0.98	1.00	0.99
FACIAL	0.96	1.00	0.94
MAQUIAGEM ROSTO	0.97	1.00	0.98

3.3 Especificidade

Na prática, quase sempre temos que escolher entre uma alta sensibilidade ou uma alta especificidade pois é impossível ter ambos. No nosso estudo, priorizaríamos ter uma alta especificidade (minimizar o FP) e podemos tolerar uma sensibilidade mais baixa (minimizar o FN). Para o modelo NB o resultado para todas as FAMILIAS ficou igual a 0%, ou seja, não foi capaz de predizer nossa classe de interesse: Modificado (y=1), sendo que em nosso conjunto de dados de treino e teste possui valores *y=0* e *y=1*. Os modelos GLM e SVM tiveram desempenhos semelhantes para todas as FAMILIAS, sendo bons modelos para prever nossa classe de interesse Modificado (y=1).

Tabela 6: E: Especificidade para os modelos propostos e seus subconjuntos.

	Modelo		
	GLM	NB	SVM
CABELO	0.66	0.00	0.67
CORPORAL	0.57	0.00	0.57
DEO COL	0.52	0.00	0.49
FACIAL	0.49	0.00	0.51
MAQUIAGEM ROSTO	0.87	0.00	0.85

3.4 Precisão

A precisão calcula a proporção do número total classificados como positivos corretamente (TP) dividido pelo número total classificados como positivo pelo modelo (FP). Quando maximizamos a precisão (mais próximo de 1), minimizamos os FP. Para os modelos GLM e SVM tivemos métricas similares entre os modelos e FAMILIAS e o modelo NB ficou com um desempenho um pouco abaixo.

Tabela 7: P: Precisão para os modelos propostos e seus subconjuntos.

	Modelo		
	GLM	NB	SVM
CABELO	0.92	0.80	0.92
CORPORAL	0.91	0.83	0.91
DEO COL	0.96	0.92	0.96
FACIAL	0.90	0.83	0.91
MAQUIAGEM ROSTO	0.97	0.81	0.97

3.5 F-Score

O desempenho de um modelo pode ser resumido por uma única pontuação que calcula a média da precisão e da sensibilidade, denominada F-Score. A maximização do F-Score maximizará a precisão e a sensibilidade ao mesmo tempo. Os três modelos obtiveram desempenhos semelhantes entre as FAMILIAS sendo que o modelo GLM obteve um leve melhor resultado.

Tabela 8: F: F-Score para os modelos propostos e seus subconjuntos.

	Modelo		
	GLM	NB	SVM
CABELO	0.93	0.89	0.93
CORPORAL	0.94	0.91	0.93
DEO COL	0.97	0.96	0.97
FACIAL	0.93	0.91	0.92
MAQUIAGEM ROSTO	0.97	0.90	0.97

3.6 Curva ROC e AUC

Sensibilidade e a especificidade são características competitivas, ou seja, é difícil aumentar a sensibilidade e a especificidade ao mesmo tempo. A curva ROC [7] é uma forma de representar a relação, entre a sensibilidade e a especificidade sendo uma demonstração bidimensional da performance do classificador. A curva ROC pode ser usada para produzir a métrica de área sobre a curva (AUC). A AUC é simplesmente a área total sob a curva ROC. Quanto maior o valor da AUC, mais eficaz é o classificador. Para os modelos GLM e SVM as áreas de todas as FAMILIAS ficaram semelhantes. Para o modelo NB, todas as FAMILIAS tiveram pior resultado comparando com os demais modelos. Dentre as FAMILIAS com melhor resultado podemos citar MAQUIAGEM ROSTO e CABELO.

Tabela 9: AUC para os modelos propostos e seus subconjuntos.

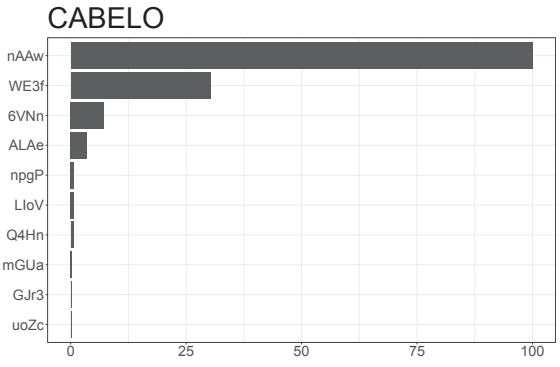
	Modelo		
	GLM	NB	SVM
CABELO	0.80	0.50	0.81
CORPORAL	0.78	0.50	0.77
DEO COL	0.71	0.50	0.71
FACIAL	0.75	0.50	0.75
MAQUIAGEM ROSTO	0.92	0.50	0.92

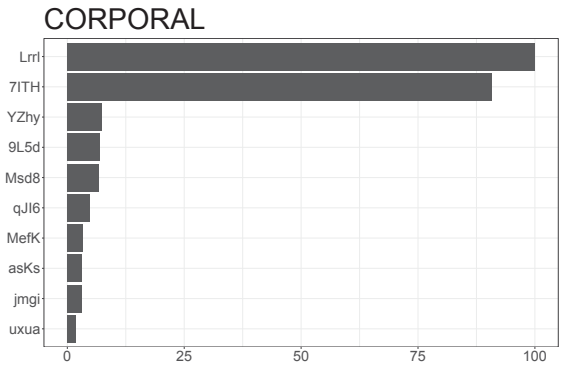
3.7 Importância variável

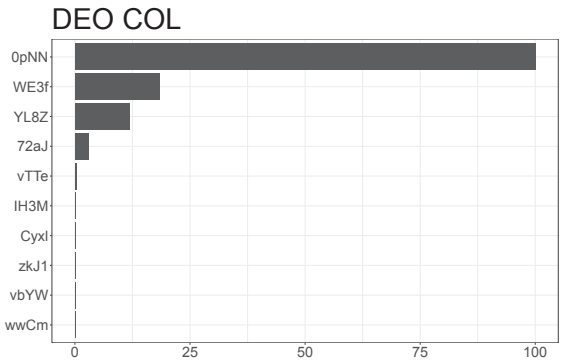
A função de avaliação de importância variável está ligada ao desempenho do modelo ser capaz de encontrar as variáveis mais importantes que contribuem mais significativamente para a predição da variável resposta(y=1). Todas as medidas de importância são dimensionadas para ter um valor máximo de 100. É utilizada a função `varImp`, do pacote `caret` dentro do R, para extrair esses dados considerando o modelo GLM com melhor desempenho. É possível notar que algumas variáveis destacam-se como relevantes para a predição pois os gráficos mostram um ligeiro decaimento da importância das variáveis em função do seu posto. Isso é importante do ponto de vista prático porque indica as variáveis de maior contribuição para a previsão de alteração. Sendo assim, pode-se estabelecer monitoramento mais rigoroso em produtos que contenham tais variáveis na composição.

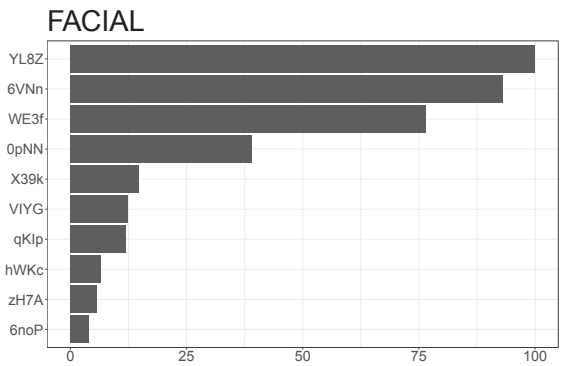
- FAMILIA CABELO, CORPORAL E DEO COL - Concentração nas duas primeiras variáveis, a partir da terceira variável temos uma baixa contribuição;
- FACIAL e MAQUIAGEM ROSTO - Concentração distribuída entre as dez primeiras variáveis.

A Figura 3 mostra que para cada FAMILIA os componenentes e sua importância são diferentes.









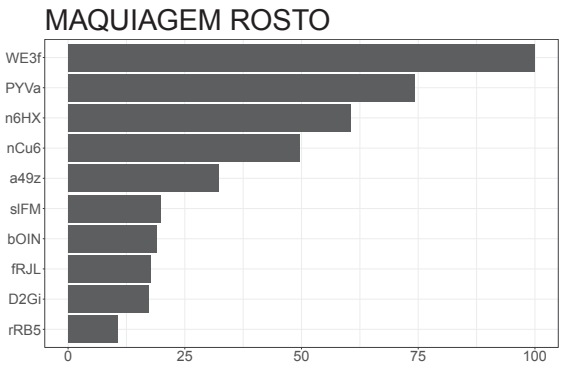


Figura 3: Dez maiores variáveis e sua importância por FAMILIA para o modelo GLM.

4 Conclusões

Observa-se que o modelo GLM obteve melhor Acurácia entre os modelos propostos para os 5 subconjuntos do atributo FAMILIA, ficando com uma média de 92% contra 91% do modelo SVM e 84% do modelo Naive Bayes. A especificidade também é uma métrica importante que deve ser considerada no estudo, sendo o modelo Glment que obteve um melhor desempenho entre os modelos testados. Outras características como dados de material de embalagem, reações adversas e *shelf life* podem ser adicionados ao modelo podendo melhorar a acurácia e a especificidade do modelo proposto. O foco do trabalho foi utilizar métodos supervisionados para predizer com um bom nível de acurácia se o estudo sofrerá alteração no t90. Os resultados obtidos neste trabalho mostram que é possível implementar um modelo de predição para estudos de estabilidade. Embora o modelo tenha sido treinado apenas para predição de estudos de estabilidade, consideramos que tais resultados podem ser expandidos para outros contextos e para as demais FAMILIAS.

Agradecimentos

Agradeço à minha esposa Gláucia Krefer, pelo apoio e compreensão durante o período do curso. Ao orientador Prof. Walmes Marques Zeviani, pelas importantes contribuições e autonomia que me concedeu durante o trabalho. Aos meus colegas de especialização e as coordenadoras dentro da diretoria de P&D Suelen Schneider e Camila Urio pela oportunidade do projeto e pelo aprendizado adquirido. E por fim, aos professores coordenadores do curso Wagner Bonat e Walmes Marques Zeviani, por viabilizarem essa oportunidade de qualificação profissional.

Referências

[1] ANVISA BRASIL. Guia de estabilidade de produto cosméticos: Series temáticas, 2004. Acesso em : 17 ago 2020.

[2] International Federation of Societies of Cosmetic Chemists. *The Fundamentals of Stability Testing*. IFSCC monograph. Micelle Press, 1992.

[3] Pete Chapman Ncr, Julian Clinton, Randy Kerber Ncr, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0. Acesso em : 17 ago 2020.

[4] Hadley Wickham and Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc., 2016.

[5] J. Grus. *Data Science do Zero: Primeiras Regras com o Python*. Alta Books, 2019.

[6] Andrew Bruce and Peter Bruce. *Estatatística Prática para Cientistas de Dados*. Alta Books, 2019.

[7] Tom Fawcett and Foster Provost. *Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. Alta Books Editora, 2018.