Metadados e flora intestinal: informações essenciais para a classificação de doenças com algoritmos de Machine Learning

1st Luan Ícaro Ferreira Santos
Engenharia de Computação
Universidade Federal do Ceará
Fortaleza, Brasil
luanicaro99@gmail.com

2st Tácio Soares Aguiar
Engenharia da Computação
Universidade Federal do Ceará
Fortaleza - CE, Brasil
tacioaguiar@gmail.com

***Abstract*—O objetivo do trabalho é conseguir classificar diversas doenças não correlatas, através de dados obtidos a partir da flora intestinal dos indivíduos. Para isso faremos o uso de diversos métodos de aprendizagem de máquina, dessa forma poderemos identificar o modelo para a classificação de uma doença específica como também, o melhor modelo para classificar um conjunto de doenças.**

I. Introdução

O uso de apredizagem estatística é essencial em várias áreas do conhecimento humano, em especial, na medicina. Devido a grande capacidade de melhorar a saúde das pessoas e, até, salvar vidas. Com essa motivação, o trabalho, usando os dados retirados do artigo: [1] Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights, foca na classificação de 13 doenças intestinais. Isso, através do uso de metadados e de informações sobre a flora intestinal das pessoas, que estão no dataset, em métodos como QDA e LDA (Linear and Quadratic Discriminant Analysis), Regressão logística, KNN (K-Nearest Neighbors) e SVM (Support Vector Machine).

II. Métodos

A. Os dados

O dataset tem duas partes principais: Os organismos e as informações pessoais dos indivíduos, no sentido de país, idade e etc. O tamanho do conjunto de dados é enorme, com 3610 linhas e 3513 colunas, então a análise exploratória mais limitada. Porém, a definição dos preditores e da saída não é complicada de ser feita. Devido a leitura do artigo [1], é fácil definir que a saída dos dados são as doenças e, por consequencia da grande quantidade de dados, as outras colunas serem os preditores. Exemplo na Fig. 1.

B. Análise Exploratória dos Dados

Para entender melhor sobre as doenças que queremos prever utilizamos esse gráfico de barras para nos informar sobre a quantidade de cada doença encontrada, os principais estados de saúde foram:

- Cancer

	country	bodysite	age	k__Archaea
0	tanzania	stool	40	0.24169
1	tanzania	stool	29	0.50621
2	tanzania	stool	8	0.30522
3	tanzania	stool	34	0.40133
4	tanzania	stool	30	0.17479
...	...	...	...	...
3605	france	stool	63	6.59835
3606	france	stool	66	0.00000
3607	france	stool	53	0.00000
3608	france	stool	63	0.17495
3609	france	stool	55	0.08997

Fig. 1. Tabela com uma amostra dos dados

- Cirrose
- Doença de Crohn
- Úlcera no colon
- Pré Diabetes
- Grande tumor Benígno no colon
- Obesidade
- Pequeno tumor Benígno no colon
- Infecção Bacteriana
- Diabetes

Dessas doenças as que serão foco de estudo nesse trabalho são: Cancer, Cirrose, Obesidade e Diabetes.

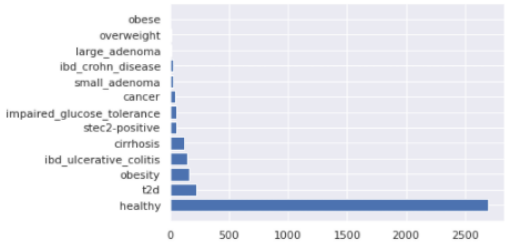

Fig. 2. Doenças

O scatterplot é um método ótimo para analisar os dados, é um gráfico de dispersão que relaciona duas variáveis e ajuda na

compreensão dos dados. No caso da Fig.3, mostra as doenças no eixo y e a idade no eixo x, mostrando sua relação. É preceptível que a maior parte dos indíviduos são saudáveis. Além disso, nota-se também que os dados são divididos de forma linear e isso é importante para definir alguns parametros e a perfomance dos modelos.

Fig. 3. Scatter Doenças x Idade

No tratamento de dados, o dataset é transformado de categórico para numérico (é melhor explicado na secção C), o que torna fácil de fazer análise de dados. Porém, é preciso um dicionário explicando qual número representa cada doença, com o intuito de ser possível compreender as imagens.

Dicionário das Doenças

- 0: 'cancer'
- 1: 'cirrhosis'
- 2: 'healthy'
- 3: 'ibd crohn disease'
- 4: 'ibd ulcerative colitis'
- 5: 'impaired glucose tolerance'
- 6: 'large adenoma'
- 7: 'obese'
- 8: 'obesity',
- 9: 'overweight',
- 10: small adenoma'
- 11: 'stec2-positive',
- 12: 't2d'

Uma boa ferramenta para entender os dados é o boxplot. Ele é um diagrama de caixa que usa os quartis dos dados, a mediana e os outliers. Na fig.4, os pontos acima da caixa, apesar de geralmente serem considerados outliers, eles não são. O fato da caixa estar localizada linha do 2, significa que a mediana e os quartis estão nos 2, ou seja, a maior parte das pessoas do dataset são saudáveis.

C. Tratamento dos dados

Como já foi elencado sobre as duas partes principais do dataset, é necessário transformar essas metades em conjunto de dados separados. Porque parte dos metadados, que são as informações pessoais, são dados categoricos, já os dados sobre os organismo vivos na flora instestinal são numericos. Então, o tratamento desses dois conjuntos se torna diferente.

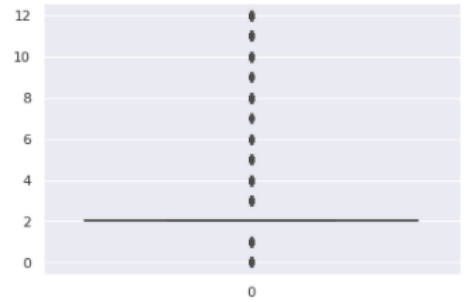

Fig. 4. Boxplot Doenças

Organismos: após a separação do conjunto original para o tratamento dos dados, é necessário descobri quais colunas desses dados precisam permanecer no dataset. Isso será determinado pela PCA (Principal Component Analyis), que será melhor explicada na subseção dedicada para ela, e das 3302 dimensões do dataset dos organismos, sobram apenas 3 colunas.

Metadados: esses dados são formados de dados numéricos e categóricos. Para limpar esses dados, é preciso separar essas informações em dois dataset. Os dados categóricos tiveram seus dados nulos transformados em 0 e o restante dos dados foram transformados em números para que possam serem aplicados no modelo. Depois os datasets são concatenados e é usado a transformação logarítmica para resolver o skewness.

D. Análise Mono e Bi Variada

Para essas análises utilizamos apenas quatro colunas dos preditores já que ficaria inviável fazer a análise de todo o dataset.

1) Mono-variada : Escolhemos os seguintes preditores:

- Country - O país em que os dados das amostras foram coletados.
- Bodysite - Parte do corpo que foi retirado a amostra.
- Age - Idade do inidivíduo que foi retirado a amostra.
- k_{Archea} - Quantidade da concentração dessa bactéria encontrada.

Todos esses preditores menos o k_{Archea} são categóricos e portanto seus dados foram transformados em números para medirmos suas métricas como, média, desvio padrão e skewness. Essas métricas foram calculadas e estão dispostas na tabela. Entre esses o k_{Archea} tem um alto valor de skewness, e mesmo após o tratamento nos preditores ele permaneceu alto devido a natureza dos dados.

Preditores	Média	Desvio Padrão	Skewness
Country	9.402	6.870	-0.264
Bodysite	14.884	7.604	-1,220
Age	32.713	17.643	1.05
$k_{Archaea}$	0.385	2.122	11.958

E. PCA

A PCA (Principal Analysis Component) é um método para reduzir a dimensionalidade do dataset para poder fazer

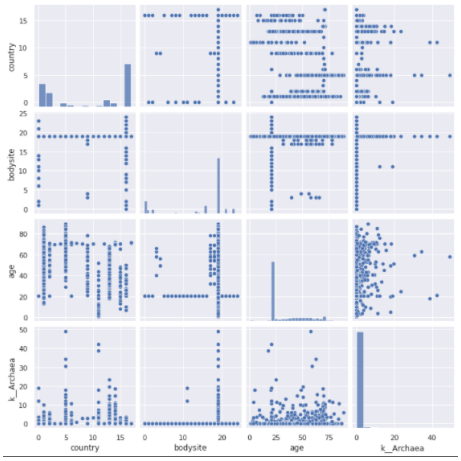

Fig. 5. Análise Bi-Variada

uma visualização dos dados, mantendo a sua variância. Essa estatística é importante porque preserva as informações e consegue criar uma representação fidedigna dos dados. A PCA seleciona as principais colunas do dataset através do cálculo dos autovetores da matriz de covariância do dataset e é selecionado quais são os principais componentes através dos maiores autovalores. Ou seja, a coluna com maior autovalor é o primeiro componente principal. A PCA dos dados é a fig. 6.

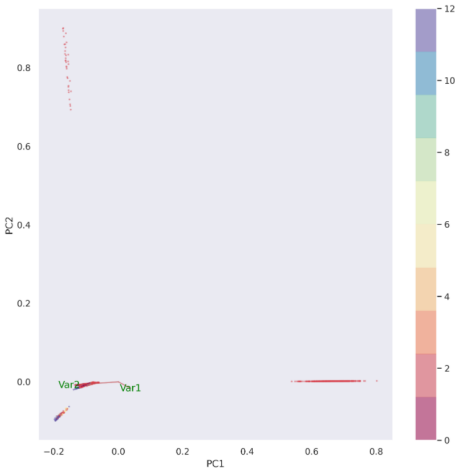

Fig. 6. PCA

F. Matriz correlação

Para diminuir o tamanho do dataset, usar uma matriz de correlação (Fig.7) para remover os dados que são altamente correlacionados é um ótimo recurso, porque aumenta o desempenho dos modelos sem prejudicar o resultado. Foram removidos as colunas com 95 % de correlação.

G. Modelos para Classificação

1) Regressão Logistica: A Regressão logistica é um método que pode ser utilizado para classificação, ele possui uma saída

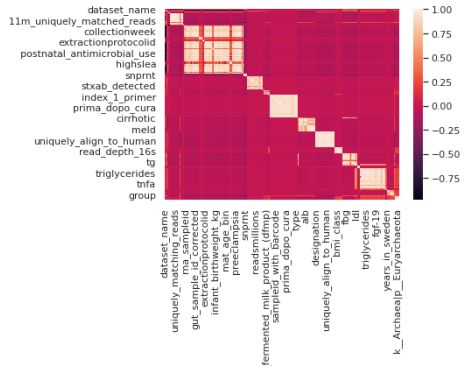

Fig. 7. Correlação entre Preditores

que será discreta e pode ter um ou mais preditores, esses podem ser discretos ou contínuos. A regressão logistica vai ajustar uma curva dependendo da probabilidade de pertencer a alguma classe. A seguir a fórmula da regressão logistica múltipla:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \quad (1)$$

- p_X - É a probabilidade da classe X
- x - É o vetor dos preditores
- β - É um parâmetro estimado a partir da máxima verossimilhança

O objetivo maior é traçar uma curva em que tenha o melhor curva que se ajuste aos dados, para assim classificar diferentes classes.

2) Análise Discriminante Linear: A Análise Discriminante Linear (LDA) é um método supervisionado que também possui o objetivo de classificação, fazendo uma redução da dimensionalidade como a PCA, com a diferença no foco em maximizar a separatibilidade entre as categorias. Uma de suas vantagens é a não existência de hiperparametros para serem ajustados. O método funciona criando linhas que maximize a distância entre as médias e minimize o ”espalhamento” de cada categoria simultanêamente. Isso pode ser representado pelo critério de Fisher a seguir:

$$\frac{(\mu_1' - (\mu_2'))^2}{\hat{s}_1^2 + \hat{s}_2^2} \quad (2)$$

- μ - Média da projeção na linha
- S - Váriancia ou o ”espalhamento” dos dados

Esse método assume que os dados de cada classe vem de uma distribuição normal, cada um com uma média e a variância comum, então utiliza esses parâmetros no classificador de Bayes. As estimativas dos parêmtros podem ser obtidos pelas seguintes equações:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)^2 \quad (4)$$

- $\hat{\mu}_k$ - Média das K-ésimas classes
- $\hat{\sigma}^2$ - Váriancia das K-ésimas classes

O ultimo parâmetro $\hat{\pi}_k$ é a probabilidade anterior de uma observação pertencer a K-ésima classe, ele é obtido da seguinte forma:

$$\hat{\pi}_k = n_k/n \quad (5)$$

Que nos levam à função do discriminante $\hat{\delta}_k(x)$ quando a quantidade de preditores for um $p > 1$:

$$\hat{\delta}_k(x) = \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k) \quad (6)$$

E para quando $p > 1$:

$$\hat{\delta}_k(x) = x^T\sum^{-1}\mu_k - \frac{1}{2}\mu_k^T\sum^{-1}\mu_k + log(\pi_k) \quad (7)$$

3) Análise Discriminante Quadrático: A Análise de Discriminante Quadrático (QDA) é um classificador que assume que as observações de cada classe é tirada da distribuição Gaussiana e plugada no estimador para os parâmetros do teorema Bayes [1]. E diferente do LDA, o QDA assume que cada classe possui sua própria matriz de covariância. Diferentemente do LDA o QDA é mais indicado quando o dataset de treino é maior pois ai a variância do classificador não é tão preocupante. Assim como na LDA o QDA utiliza os estimadores do $\sum_k$, μ_k e π_k em sua equação abaixo:

$$\delta_k(x) = -\frac{1}{2}x^T\sum_k^{-1}x + x^T\sum_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\sum_k^{-1}\mu_k - \frac{1}{2}log\left|\sum_k\right| + log(\pi_k) \quad (8)$$

Diferente da equação do LDA a quantidade de x aparece como uma função quadrática.

4) K Vizinhos mais próximos: Esse método também é utilizado na classificação, onde dado um conjunto de dados de diferentes classes, é calculado a distancia entre os indivíduos no espaço, então supõe-se que, por similaridade, os dados que estiverem mais próximos pertencem à mesma classe e os que estiverem mais longes são de outra classe. Esse método utiliza um parâmetro o K, que será a quantidade de vizinhos a se considerar quando aplicado o método. É possível obter diferentes resultados utilizando diferentes K, portanto normalmente são testados várias vezes diferentes valores nesse parâmetro para então obter a melhor perfórmance do modelo.

5) Máquina de Vetor de Suporte: Esse método de classificação constrói de linhas a hiperplanos em um espaço n-dimensional considerando o dataset e os pontos observados de cada classe nele, para então ser separado. A margem desses espaços divisores é um ponto muito importante nesse método, onde uma margem maior reduz a variância do modelo e uma margem maior aumenta, ambos os casos o modelo estaria enviesado. Portanto esxiste dois métodos adotados para a escolha do tamanho da margem, o Soft Margin e Hard Margin, dependendo do dataset. É idicado utilizar Hard Margin no dataset linearmente separáveis, e o Soft Margin performa melhor nos datasets linearmente inseparáveis, podendo trazer um tratamento melhor com os outliers. O SVM utiliza Kernel, que nada mais é que uma função que quantifica a similaridade de duas observações. Existem Kernel do tipo linear, polinomial, radial e a maior vantagem de seu uso é computacional, podendo trabalhar até com espaços vetoriais virtualmente infinitos.

H. ROC

A Curva Característica de Operação do Receptor, ou sua sigla em inglês ROC, é um gráfico que permite de uma forma simples sumarizar vários limiares dos um ou mais modelos e assim torna mais fácil a comparação entre eles.

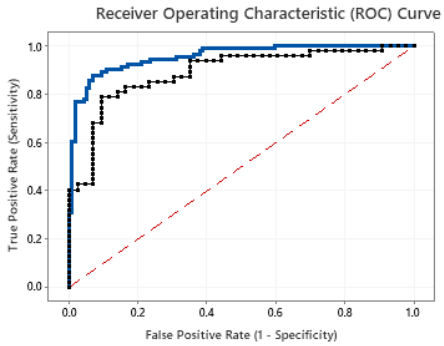

Fig. 8. ROC exemple

A figura 1 é um exemplo genérico da curva, em que o eixo Y representa os Verdadeiros Positivos e o eixo X os Falsos Positivos. A linha pontilhada vermelha basicamente mostra se o modelo com aquele threshhold está mais acertivo ou não. Portanto quanto mais inclinada ao canto superior esquerdo do gráfico estiver a curva, melhor será aquele modelo. Uma forma mais precisa de comparar os modelos nesse gráfico é calculando a área sob a curva. Assim, os modelos que tiverem a maior área são em geral melhor.

III. Resultados

Para entender as matrizes de confusão é necessário usar o Dicionário 1 e para a ROC curve o Dicionário das Doenças ROC

Dicionário Doenças ROC

- class 0: 'cancer'
- class 1: 'cirrhosis'
- class 2: 'healthy'
- class 3 : 'obesity'
- class 4: 't2d'

A. Matriz de confusão

As matrizes de confusão mostram quantos verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos tem o resultado do modelo. Isso é calculado através da comparação entre o resultado previsto e qual o resultado verdeiro. No trabalho, é usado um heatmap com as % de verdadeiros positivos e falsos positivos.

B. ROC Curve dos modelos

A ROC curve dos modelos mostra que as pessoas saudáveis são as com maior quantidade erro, mas isso porque, os modelos tendem a errar para os indivíduos saudáveis. Depois disso, a diabetes é a doença que os modelos erram mais, sendo a segunda curva menos acentuada e a segunda de menor área.

C. Regressão Logistica

Com a aplicação da fórmula 1, o modelo faz os cálculos da classificação multinomial, determinando valores entre 0 e 12. O modelo se saiu bem porque usa limites de decisão lineares para determinar qual é o resultado. O que para esses dados é determinante para ter um bom resultado, porque os dados são separados de forma linear, como mostra na fig.3.

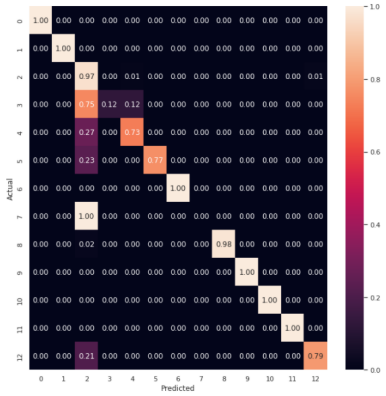

Fig. 9. Matriz de Confusão - Regressão Logistica

Os resultados da matriz de confusão, Fig.9, mostra que a maioria dos erros foi porque o modelo confundil as doenças com pessoas saudáveis, colocando elas dentro do limite de decisão errado.

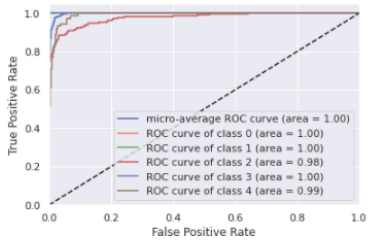

Fig. 10. ROC - Regressão Logistica

D. LDA

A análise discriminante linear também é um método que separa os modelos de forma linear, o que se adequa perfeitamente aos dados e, por isso, teve um ótimo resultado.

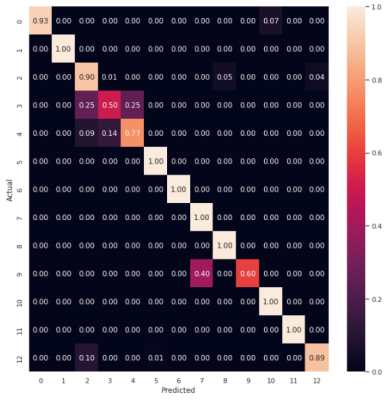

Fig. 11. Matriz de Confusão - LDA

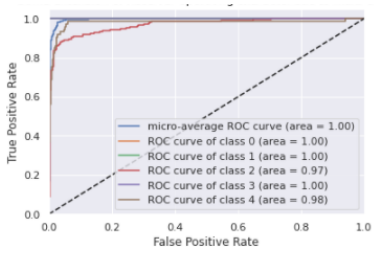

Fig. 12. ROC - LDA

E. QDA

Devido a variáveis colineares e aos dados serem separados linearmente, a QDA teve dificuldade de identificar certas doenças. Porque quando o cálculo é feito para determinar a classificação da doença, os valores colineares fazem que o modelo erre a conta e, o método, também, não consegue se adequar ao perfil linear dos dados. Por exemplo, na Fig.13 as doenças de número 0, 6 e 10 tiveram resultados na mesma coluna da matriz de confusão devido as colunas serem colineares.

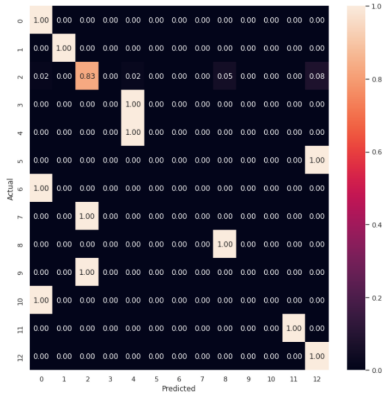

Fig. 13. Matriz de Confusão - QDA

F. KNN

Como o esse modelo não trabalha com separações lineares, apresenta um resultado mediano para ruim na maioria das

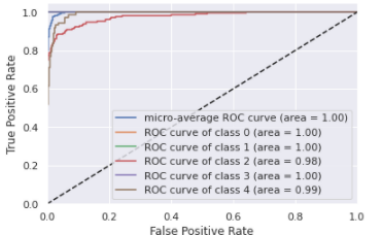

Fig. 14. ROC - QDA

classes de doenças como mostrado na Fig 15, com maior erro em classificar a classe '2', saudável.Pois devido ao maior número de dados nessa classe, não foi possível achar um k bom suficiente para conseguir separar essas classes, mesmo usando um método de cross-validation para achar os melhores parâmetros.

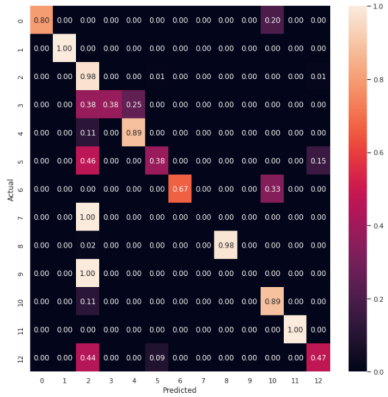

Fig. 15. Matriz de Confusão - KNN

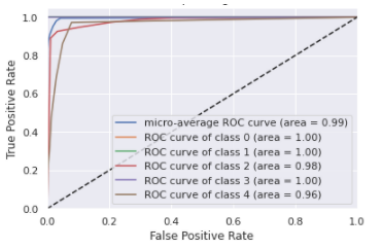

Fig. 16. ROC - KNN

G. SVM

O SVM é um modelo muito poderoso devido ao kernel. Ele pode mudar completamente como os vetores irão delimitar os dados e, no caso desse dataset, aplicando um núcleo linear faz que o modelo tenha o melhor desempenho para quase todas as doenças, como mostra na Fig.17.

H. Resultados dos Modelos

Observando as matrizes de confusão e as curvas ROC de cada modelo, podemos inferir que cada caso tem sua importância e deve ser analisada de forma separada. Olhando

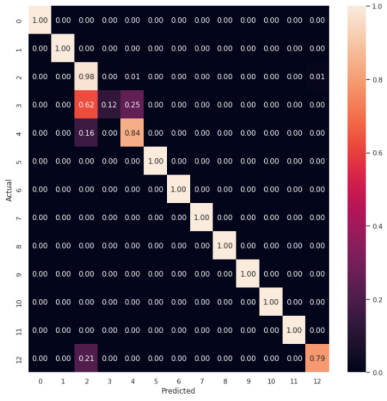

Fig. 17. Matriz de Confusão - SVM

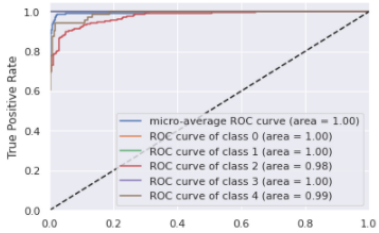

Fig. 18. ROC - SVM

em um contexto geral, contando com todas as doenças, temos o modelo que obteve o melhor desempenho, Suport Vector Machine, com uma acurácia de 95%. Esse modelo poderia ser bem aproveitado quando não se tem uma ideia exata da doença que está sendo investigada, pois existem altas chances de ele acertar com uma boa acurácia. Mas, ele falha em algumas doenças pontuais, e é ai que vale a pena checar os outros modelos. Caso tenha alguma doença específica que está sendo investigada que o SVM não possui uma taxa de acerto boa o suficiente, como a '12: t2d' - Diabetes, um modelo como o QDA, possui uma alta taxa de acerto dessa doença, portanto ele seria o mais indicado mesmo dentre todos os outros ele sendo o modelo com a acurácia geral mais baixa. Então para cada caso, vale a pena observar qual modelo está mais apto a performar essas análises.

Modelo	Acurácia
LDA	90%
QDA	83%
Regressão logística	94%
SVM	95%
KNN	91%

REFERENCES

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.
[2] JKuhn, M., Johnson, K. (2018). Applied predictive modeling. Springer.
[3] Pasolli E, Truong DT, Malik F, Waldron L, Segata N (2016) Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol 12(7): e1004977.