INF0615 – Aprendizado de Máquina Supervisionado

Trabalho 2 - Regressão Logistica

Rodolfo Dalla Costa

Nicole Nogueira

9/11/2021

# Introdução

O sistema imunológico humano é o sistema responsável por proteger o corpo de antígenos como vírus e bactérias. A produção de glóbulos brancos, nome dado às células que compoem o sistema imunologico, é originada por algumas cadeias proteicas presentes no antigeno. Desse modo, propoe-se utilizar um modelo de regressao logistica com o objetivo, a partir de determinadas caracteristicas de uma cadeia proteica, a mesma pode gerar uma resposta do sistema imunilogico.

# Banco de dados

Na pode ser observado os dados que foram utilizados para o desenvolvimento do trabalho. Nota-se que todos praticamente sao dados numericos, e a coluna target é a coluna resultado. A base foi dividida em 3 partes da base e uma outra base externa para testes, sendo portanto, 1 parte de treino, 1 de validaçao, 1 de teste e 1 de teste sobre o virus SARS. Cada uma contem respectivamente 9204, 2303, 2878 e 520 linhas e um total de 11 colunas (como observado acima).

# Análise Descritiva

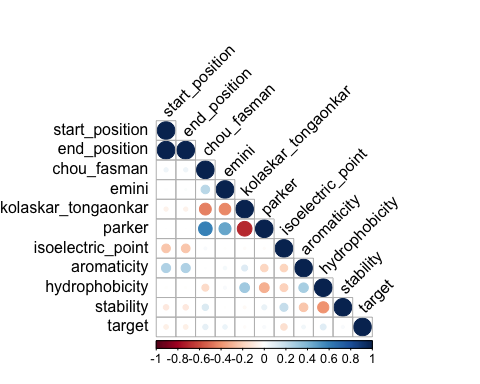
## [1] "Dados Faltantes no treino: FALSE"

## [1] "Dados Faltantes na validação: FALSE"

## [1] "Dados Faltantes no teste: FALSE"

## [1] "Dados Faltantes no SARS: FALSE"

Como pode ser observado, a base nao possui nenhum dado faltante para nenhuma das partes.



Correlações 2 a 2 das variáveis.

Dado o mapa de correlacoes da Figura , observa-se que a variavel target nao possui correlacoes com um valor evidentemente alto, sendo notavel uma correlacao inversa com a variavel target com a variavel isoeletric\_point e uma correlacao inversa mais baixa com start\_position e end\_position. Em contrapartida nota-se um correlacao positiva, porem mais fraca, com as variaveis chou\_fasm, hydrophobicity, emini e aromaticity.

# Metodologia

Apos a etapda de inspecao dos dados, foi realizada a normalizacao utilizando o metodo Z-Norm. Em seguida, a partir dos dados nao balanceados foi gerada uma Baseline considerando todas as variaveis num polinomio de grau 1. Apos isso, foram aplicadas 3 tecnicas de balanceamento: SMOTE, undersample e a ponderada por pesos; dentre elas o balanceamento ponderado gerou melhores resultados. A partir disso, uma serie de hipoteses foram testadas para gerar o modelo: polinomios de 1 a 12, combinacoes considerando as variaveis de maior correlacao e a propria baseline.

# Resultados e Conclusão

##   
## 0 1   
## 6709 2495

##   
## 0 1   
## 0.729 0.271