UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

TRABALHO DE APRENDIZAGEM DE MÁQUINA
CLASSIFICAÇÃO E REGRESSÃO SOBRE DADOS EM SAÚDE FETAL

DANIELLE LIMEIRA SILVA
DANILO IZAÍAS DE MACÊDO
JOÁS DE BRITO FERREIRA FILHO
MARINA RODRIGUES DE OLIVEIRA

João Pessoa, Novembro de 2020

1 INTRODUÇÃO

Questões como pobreza, distância ao acesso à saúde, falta de informação, serviços inadequados e outras, são algumas das razões que colaboram para a continuidade do alto número de mortalidade infantil e materna. Segundo a Organização das Nações Unidas (ONU) (2019), a cada 11 segundos uma grávida ou um recém-nascido morrem e a Organização Pan-Americana da Saúde (OPAS) (2018) afirma que 830 mulheres morrem todos os dias por complicações relacionadas à gravidez ou ao parto em todo o mundo, sendo 99% em países em desenvolvimento.

A realidade da mortalidade infantil e materna é muito desigual entre os países, os maiores números se concentram na África Subsaariana (ONU, 2019). Em 2018, uma em cada 13 crianças faleceram antes dos cinco anos na África Subsaariana, número 15 vezes maior comparado à Europa (uma criança a cada 196 nascidos vivos) (ONU, 2019). O número de mortalidade materna e infantil na África Subsaariana e no sul da Ásia representaram 80% do total (ONU,2019).

Segundo a ONU (2019), mesmo com números tão alarmantes, tem sido registrado progressos. Entre 1990 e 2018 o número de mortes de crianças com menos de 15 anos reduziu 56% (saindo de 14,2 milhões para 6,2 milhões) e a mortalidade materna reduziu 38% entre 2000 e 2017, sendo a maior redução no sul da Ásia (60%) (ONU, 2019). A ONU espera erradicar, até 2030, as mortes evitáveis de recém-nascidos e de crianças menores de 5 anos, com todos os países tendo o mesmo objetivo de reduzir a mortalidade neonatal para pelo menos 12 a cada 1000 nascidos vivos e 25 a cada 1000 nascidos vivos para a mortalidade de 5 anos (ONU, 2019).

Os meios para prevenção já é bem conhecido e engloba o acesso à saúde, com a realização de consultas pré-natias e outro cuidados (OPAS, 2018). Porém, nem todas as mulheres conseguem ser assistidas por diversas razões, o que torna a disponibilidade de meios simples e de baixo custo tão importantes. Os Cardiogramas são uma opção que com esse caráter que peritem aos profissionais de saúde trabalharem na prevenção infantil e materna (KAGGLE, 2020). O equipamento envia pulsos de ultrassom e lêem as respostas, permitindo a análise da frequência cardíaca fetal (FCF), movimentos fetais, contrações uterinas e mais, sendo possível identificar se há alguma possível patologia no feto (KAGGLE, 2020).

Tendo em vista essa realidade, o presente trabalho tem como objetivo desenvolver métodos de aprendizagem de máquina para a identificação da situação de saúde do feto, através do conjunto de dados *Fetal Health Classification*, disponibilizado no site *Kaggle*. Esses são dados obtidos de Cardiogramas, ao todo são 2.126 registros, analisados e classificados por três obstetricistas especialistas da seguinte forma: Normal, Suspeito ou Patológico. Para isso, serão utilizas as metodologias de classificação e regressão.

Este trabalho está dividido em seis etapas a partir da introdução. O capítulo dois é destinado à base de dados, o capítulo três enuncia a análise exploratória da base utilizada, o capítulo quatro aborda a metodologia e, por fim, nos capítulos cinco e seis são apresentados os resultados e a conclusão, respectivamente.

2 BASE DE DADOS

Neste trabalho foram utilizados os dados de 2.126 registros de cardiogramas (CTG), disponibilizados na base de dados *Fetal Health Classification*, pelo site *Kaggle*, o qual não informou a origem dos dados. A base é composta por 22 variáveis resultantes deste exame, as quais são utilizadas por profissionais como uma das formas de descrever a situação de saúde do feto.

Segundo Moura et. al. (1992), no CTG, pelo sistema Porto de análise, os algoritmos de processamento são baseados em histogramas e nos pontos de sinais como máximos, mínimos, cruzamento de determinados limares e etc. Do CTG são extraídas as variáveis:

- Linha Base dos batimentos cardíacos fetais (FHR), que são medidos em batimentos por minuto. A linha base corresponde ao valor mais frequente dos FHR na ausência de contrações uterinas e de movimentos feitas. Sendo assim, é estimada pelo máximo do histograma do FHR (MOURA et. al, 1992);

- Acelerações e Desacelerações (leve, severa ou prolongada) do FHR, os quais são dados pelo desvio da linha base de pelo menos 15 batimentos por minuto, durante pelo menos 15 segundos (MOURA et. al, 1992);

- Detecção de contrações uterinas, medidas em mm Hg. É a detecção de picos e cruzamentos por valor de referência, estes ocorrem quando a amplitude do sinal de contração uterina é maior do que 15 mm de Hg, em um intervalo de 45 a 90 segundos (MOURA et. al, 1992);

- Variabilidade do FHR a longo prazo, a qual é a amplitude entre o máximo e mínimo do FHR, em valores absolutos, em intervalos de meio minuto (MOURA et. al, 1992);

- Variabilidade do FHR a curto prazo, a qual é a amplitude entre o máximo e mínimo do FHR, em valores absolutos, em intervalos de dois segundos (MOURA et. al, 1992).

A base *Fetal Health Classification* também apresenta outras variáveis, que são: valor médio da variabilidade do FHR a curto prazo; valor médio da variabilidade do FHR a longo prazo; porcentagem de tempo com variabilidade anormal de longo prazo e; os histogramas do FHR com relação à amplitude, mínimo, máximo, número de picos, número de zeros, moda, média, mediana, variância e tendência. A última variável presente na base é a Saúde Fetal, a qual corresponde a saúde do feto e que será utilizada como variável resposta nos métodos de aprendizagem de máquina. A saúde fetal foi classificada por três obstetricistas especialistas da seguinte forma: Normal, Suspeito ou Patológico.

3 ANÁLISE EXPLORATÓRIA

Inicialmente foi realizado um estudo na base de dados em questão com a finalidade de identificar anomalias distributivas, redundâncias nas informações e estatísticas básicas que permitem discriminar o comportamento de similaridade e variabilidade entre as observações.

Como resultado, não foram encontrados dados faltantes ou repetidos, porém observou-se que as classes da variável Saúde Fetal estavam desbalanceadas, 77,85% da base correspondia às informações da classe normal, 13,88% da classe suspeito e 8,28% da classe patológico. Tendo em vista essa dificuldade, foi realizado o balanceamento da base por *oversampling*. Primeiramente calculou-se a proporção entre as classes, sendo constatado que a proporção da classe normal pela classe suspeito foi de 5,61 e pela classe patológico foi de 9,40. Logo, a classe suspeito foi replicada cinco vezes e a patológico nove vezes. Por fim, foi adicionada uma amostra às classes suspeito e patológico com tamanhos iguais às diferenças restantes entre o tamanho da classe normal e as outras duas, obtendo uma base com 1.655 observações para cada classe da variável saúde fetal.

Com a base balanceada, buscou-se analisar a necessidade de utilização de todas as 22 variáveis, ou se seria possível realizar uma redução de dimensionalidade, permitindo estudar as variáveis que possuem maior impacto na variabilidade da base. Então, aplicou-se a análise de componentes principais. Calculou-se os autovetores, com base na correlação entre todas as variáveis, e foi calculada a proporção de variabilidade, identificando que aproximadamente 99,32% da variabilidade da base se concentrava nas 13 primeiras variáveis, com exceção da variável saúde fetal. Sendo assim, construiu-se uma nova base contento 14 variáveis, as quais são: movimento fetal, porcentagem de tempo com variabilidade anormal de longo prazo (PTVALP), valor médio da variabilidade de curto prazo (VMVCP), variabilidade anormal de curto prazo (VACP), acelerações, desacelerações leves, desacelerações prolongadas, desacelerações severas, histograma da mediana de FHR, histograma da moda de FHR, histograma do mínimo de FHR, número de zeros do histograma, contrações uterinas e saúde fetal. Em seguida foi realizada uma breve análise descritiva sobre a base, como mostra a Tabela 1.

Tabela 1 - Análise Descritiva

Variáveis	Mínimo	Máximo	Média	Mediana	Amplitude	Desvio Padrão	Variância
Mov. Fetal	0	0.481	0.0140	0	0.481	0.0604	0.00365152
Hist. Mediana	77	186	134.995972	139	109	18.6272002	346.972587
Desa. Prolong.	0	0.005	0.00047553	0	0.005	0.001019183120	1.03873e-06
Hist. Mín	50	159	96.0682779	99	109	33.314950841812	1109.885949
PTVALP	0	91	18.9319235	4	109	33.314950841812	1109.8859496
Desa. Seve	0	0.001	1.20846e-05	0	91	25.789620989	665.10455079
Acele.	0	0.019	0.0015	0	0.019	0.0029	8.8209854e-06
Hist. Moda	60	187	133.079154	139	127	22.6523132375	513.12729501
VMVCP	0.2	7	1.21697885	0.9	6.8	1.0104229855	1.02095460974
VACP	12	87	56.3073515	61	75	17.1178487464	293.02074570
Desa. Leves	0	0.015	0.00204693	0	0.015	0.00329865995	1.0881157e-05
Hist. N. Zeros	0	10	0.30956697	0	10	0.730535161009	0.533681621
Contra. Uteri.	0	0.015	0.00365378	0.003	0.015	0.003193965959	1.02014185e-05

A Tabela 2 nos permite observar de maneira precisa a correlação que existe entre todas as variáveis, de modo que podemos perceber quais variáveis estão mais correlacionadas com a variável de saída, o que nos possibilita efetuar uma seleção de variáveis minimalista, mas ainda assim bem embasada por indicadores estatísticos. É preciso ressaltar o significado de "mais correlacionadas"quando tratamos de Coeficientes de Pearson. Um coeficiente negativo também pode indicar uma correlação forte, neste caso devemos sempre

nos atentar ao módulo dos coeficientes, o que indica a magnitude da correlação, desta forma o sinal indica a o sentido desta correlação, de maneira análoga a grandezas direta e inversamente proporcionais.

Assim, analisando os resultados, é possível observar que a correlação entre as variáveis e suas correlações com a variável respostas (saúde fetal), em geral, não são fortes, com exceção da correlação entre as variáveis de histogramas. Além disso, a influência entre as variávies é negativa, ou seja, inversa.

Por fim, houve a necessidade de aplicar a normalização dos dados, tendo em vista que as variáveis possuíam escalas diferentes. Para isso, aplicou-se a toda base a transformação que retorna valores entre zero e um.

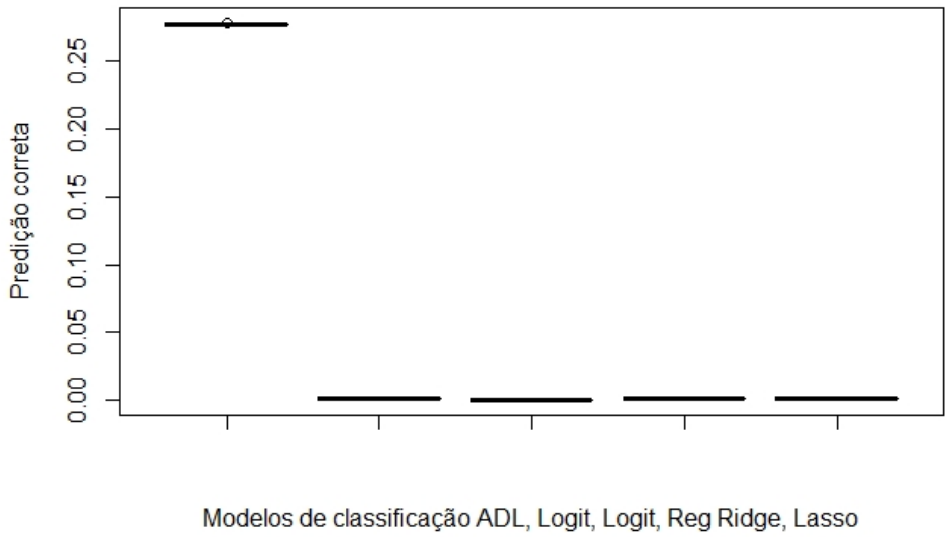
Tabela 2 - Matriz de Correlação Linear de Pearson

Variáveis	Mov. Fetal	Hist. Mediana	Desa. Prolong.	Hist. Mín	PTVALP	Desa. Seve.	Acele.	Hist. Moda	VMVCP	VACP	Desa. Leves	Saúde Fetal	Hist. N. Zeros	Contra. Uteri.
Mov. Fetal	1.0000	-0.1412	0.3440	-0.1593	-0.1193	-0.0245	-0.0092	-0.1357	0.1152	-0.1649	0.0149	0.1224	-0.0037	-0.0774
Hist. Mediana	-0.1412	1.0000	-0.6407	0.6029	0.3301	-0.2482	0.1512	0.9229	-0.5495	0.0489	-0.5771	-0.4162	-0.1483	-0.2500
Desa. Prolong.	0.3440	-0.6407	1.0000	-0.4771	-0.3366	-0.0317	-0.1059	-0.5990	0.4995	-0.0362	0.3199	0.4921	0.1126	0.1920
Hist. Mín	-0.1593	0.6029	-0.4771	1.0000	0.5323	-0.1333	-0.1430	0.5535	-0.7031	0.3114	-0.6123	-0.0896	-0.3387	-0.3285
PTVALP	-0.1193	0.3301	-0.3366	0.5323	1.0000	-0.0812	-0.3442	0.3098	-0.5798	0.4981	-0.4076	0.2781	-0.2106	-0.4739
Desa. Seve.	-0.0245	-0.2482	-0.0317	-0.1333	-0.0812	1.0000	-0.0566	-0.2940	0.0771	0.0297	0.2158	0.1309	0.1020	0.0206
Acele.	-0.0092	0.1512	-0.1059	-0.1430	-0.3442	-0.0566	1.0000	0.1306	0.1933	-0.3999	-0.0455	-0.4942	0.0057	0.1610
Hist. Moda	-0.1357	0.9229	-0.5990	0.5535	0.3098	-0.2940	0.1306	1.0000	-0.5110	0.0116	-0.5751	-0.4291	-0.1644	-0.2224
VMVCP	0.1152	-0.5495	0.4995	-0.7031	-0.5798	0.0771	0.1933	-0.5110	1.0000	-0.4505	0.6361	0.0616	0.3473	0.5229
VACP	-0.1649	0.0489	-0.0362	0.3114	0.4981	0.0297	-0.3999	0.0116	-0.4505	1.0000	-0.1291	0.5241	-0.2154	-0.3628
Desa. Leves	0.0149	-0.5771	0.3199	-0.6123	-0.4076	0.2158	-0.0455	-0.5751	0.6361	-0.1291	1.0000	0.2139	0.3257	0.4180
Saúde Fetal	0.1224	-0.4162	0.4921	-0.0896	0.2781	0.1309	-0.4942	-0.4291	0.0616	0.5241	0.2139	1.0000	0.0068	-0.1260
Hist. N. Zeros	-0.0037	-0.1483	0.1126	-0.3387	-0.2106	0.1020	0.0057	-0.1644	0.3473	-0.2154	0.3257	0.0068	1.0000	0.1713
Contra. Uteri.	-0.0774	-0.2500	0.1920	-0.3285	-0.4739	0.0206	0.1610	-0.2224	0.5229	-0.3628	0.4180	-0.1260	0.1713	1.0000

4 RESULTADOS

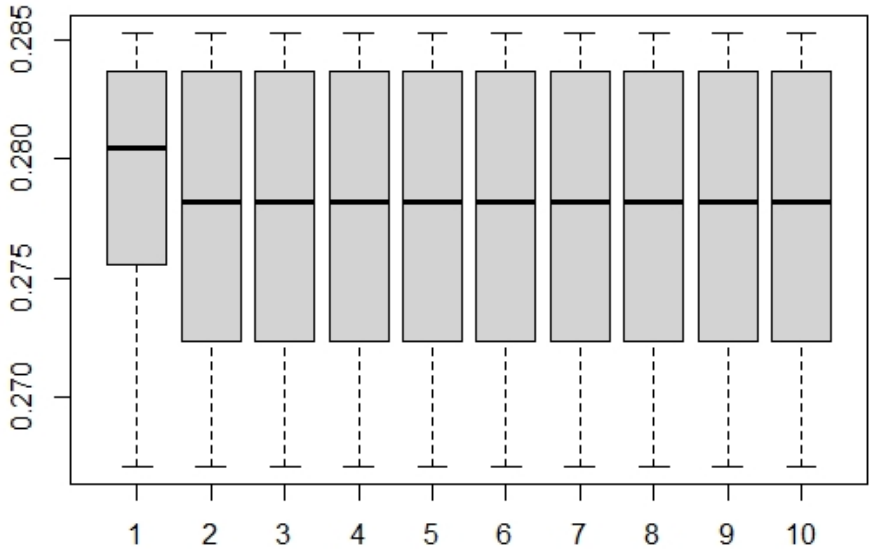
Métodos usados para classificação foram a regressão logística, Analise de discriminante Linear, regressão Ridge, Lasso, Elasticnet.

Figura 1: Boxplot com o desempenho dos modelos de Analise de discriminante Linear, regressão logística fetos com saude normal, regressão logistica fetos com saude suspeita, regressão Ridge, Lasso



Observando o boxplot é possível ver o modelo de classificação que todos os modelos têm variação baixa, com ligação logit para fetos com saúde normal e logit com saúde suspeita, regressão Ridge e regressão Lasso com péssimos desempenhos. Os modelo de classificação com desempenhos ruins foram todos com melhor nível de médio de predição correta sendo 0.2774496, não chegando nem a 30% de acerto.

Figura 2: Boxplot com o desempenho do modelo de Analise de discriminante Linear nos 10 folds da validação cruzada



O Primeiro modelo de classificação sendo Analise de Discriminante Linear teve uma desempenho de predição correta de cerca de 27,74%, o melhor desempenho dentre todos os modelos avaliados.

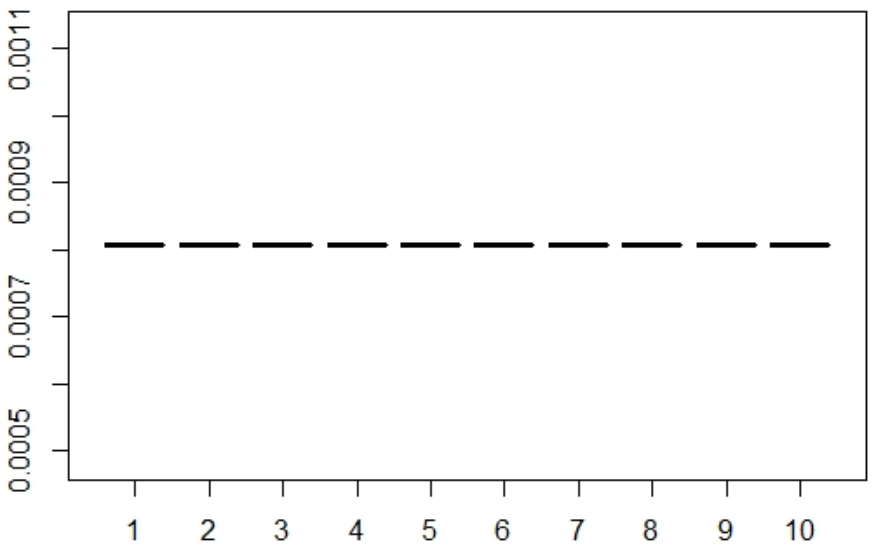
O segundo modelo de classificação sendo de regressão logística, observando o desempenho de predição correta podendo observar pelo boxplot foi um péssimo modelo para predição não chegando a 1% de acerto em média.

O terceiro modelo de classificação sendo de regressão logística para fetos com suspeita de fraude, teve um desempenho tão ruim quanto o modelo de regressão logística para fetos com saúde normal observando o desempenho de predição correta podendo observar pelo boxplot foi um péssimo modelo para predição não chegando a 1% de acerto em média, porém com maior variabilidade.

O Quarto modelo de classificação sendo de regressão ridge, teve um desempenho tão ruim quanto os modelos de regressão logística para fetos com saúde normal e com saúde suspeita observando o desempenho de predição correta podendo observar pelo boxplot não foi uma boa escolha com a predição não chegando a 1% de acerto em média, mas obteve variação baixa.

O quinto modelo de classificação sendo de regressão ridge, teve um desempenho tão ruim quanto os modelos de regressão logística para fetos com saúde normal e com saúde suspeita observando o desempenho de predição correta podendo observar pelo boxplot não foi uma boa escolha com a predição não chegando a 1% de acerto em média, mas obteve variação baixa.

Figura 3: Boxplot com o desempenho do modelo de regressão Logistica para fetos com saúde normal nos 10 folds da validação cruzada



5 CONCLUSÃO

Em geral o resultado não foi como esperado, entre os métodos utilizados o melhor método foi o análise de discriminante regressão logística em média tendo uma predição correta de 0.2774496 e os demais foram os regressão Ridge, Lasso e Elasticnet. Concluindo que o melhor modelo de classificação foi Análise de Discriminante Lienar.

6 REFERÊNCIA

https://www.paho.org/bra/index.php?option=com$_{c}ontentview = articleid = 5741 : folha{-}informativa{-}mortalidade - maternaItemid = 820$

https://news.un.org/pt/story/2019/09/1687532

https://www.msf.org.br/o-que-fazemos/atividades-medicas/saude-materna

https://www.kaggle.com/andrewmvd/fetal-health-classification

Figura 4: Boxplot com o desempenho do modelo de regressão Logistica fetos com saude Suspeita nos 10 folds da validação cruzada

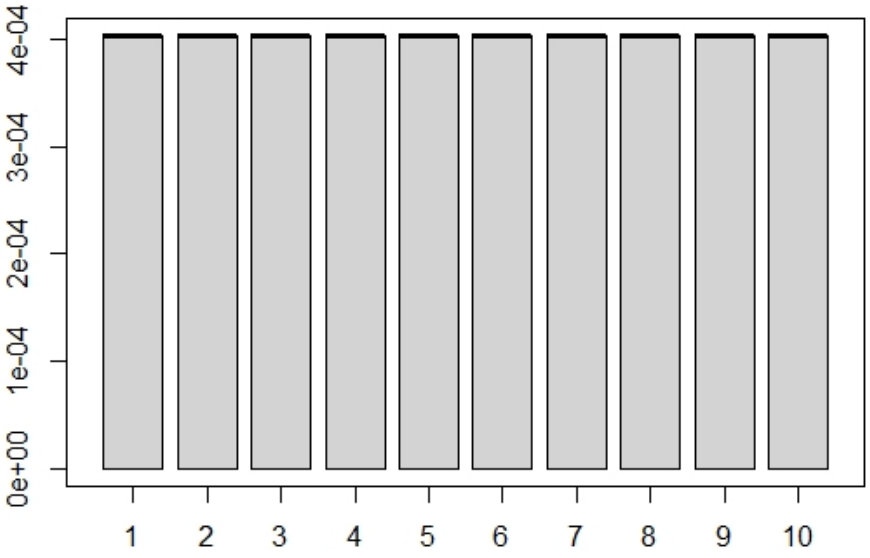


Figura 5: Boxplot com o desempenho do modelo de regressão Ridge nos 10 folds da validação cruzada

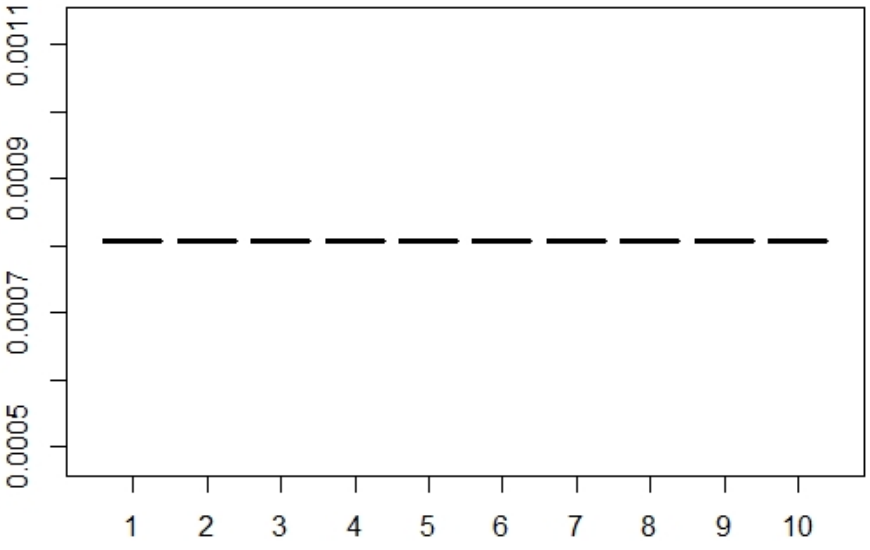


Figura 6: Boxplot com o desempenho do modelo de regressão Lasso nos 10 folds da validação cruzada